Issues in Developing a Resource-Based Relative Value Scale for Physician Work

James P. Kahan, Sally C. Morton,
Gerald F. Kominski, Hilary H. Farris,
Arthur J. Donovan, David L. Bryant

This report is made pursuant to Cooperative Agreement 99-C-98489/9-07 between the Health Care Financing Administration, U.S. Department of Health and Human Services and RAND. The cooperative agreement was competitively awarded. The amount charged to the Health Care Financing Administration for the work resulting in this report (inclusive of the amounts so charged for any prior reports submitted under this contract task) is $109,702 (100% financed with federal money; no nongovernmental sources used).

RAND

The research described in this report was supported by the Health Care Financing Administration, U.S. Department of Health and Human Services, Cooperative Agreement 99-C-98489/9-07.

ISBN: 0-8330-1259-2

Published 1992 by RAND
1700 Main Street, P.O. Box 2138, Santa Monica, CA 90407-2138

R-4130-HCFA

Issues in Developing a Resource-Based Relative Value Scale for Physician Work

James P. Kahan, Sally C. Morton,
Gerald G. Kominski, Hilary H. Farris,
Arthur J. Donovan, David L. Bryant

Supported by the
Health Care Financing Administration,
U.S. Department of Health and Human Services

RAND

RAND/UCLA/Harvard
Center for Health Care
Financing Policy Research

PREFACE

This report examines selected issues regarding the measurement and calculation of the resource-based relative values for physician work to be used in the Medicare Fee Schedule. Of particular interest are issues central to the obtaining of survey data from physicians and calculations used to transform those data to work values. Different methods of data collection and analysis have consequences in terms of the distribution of Medicare payments for different services, which must be taken into consideration in revising the system as it evolves into a "steady state." Only by understanding these consequences can the Health Care Financing Administration (HCFA) maintain a steady-state Medicare Fee Schedule that provides a stable policy for the medical community and the public yet is responsive to changes in the economics and technology of medical practice. The results reported here should be useful to HCFA in choosing a general method for adjusting work values as well as in making other medical policy choices.

The analyses reported here were performed within the RAND/UCLA/Harvard Center for Health Care Financing Policy Research.

SUMMARY

PHYSICIAN WORK VALUES FOR THE MEDICARE FEE SCHEDULE

The Medicare Fee Schedule (MFS), which took effect as mandated by Congress in January 1992, replaces the system of customary, prevailing, and reasonable charges used by the Health Care Financing Administration (HCFA) for physician payment with a system of payment based on relative value units (RVUs). The RVU for a physician's service comprises three elements: (1) the relative value for physician work (RVW); (2) practice expenses, i.e., overhead, excluding malpractice expenses; and (3) malpractice expenses. Under the MFS, payment for a service involves adjusting each of these three elements of the RVU by a separate geographical factor, then summing to produce a value for the service. This RVU is then multiplied by a national conversion factor to yield a dollar amount for payment.

The major innovation of the MFS—and the focus of this report—is the measurement of RVWs. Estimates of physician work under the new fee schedule are based on the Resource-Based Relative Value Scale (RBRVS). That scale was constructed by a team of researchers at the Harvard School of Public Health, headed by William C. Hsiao, Ph.D. As the RBRVS has evolved, it has been extensively commented upon and criticized and numerous revisions have been recommended. However, the estimates of physician work contained in the MFS rules issued to take effect in January 1992 are based primarily on the Harvard study.

The focus of the project reported on here was to examine alternative ways to: (1) obtain survey data on the amount of work performed by physicians in different specialties, and (2) transform those data to a common scale of relative work values. The work reported on here was largely conducted in 1991, before any public release of the results of Phase III of the Harvard study or the HCFA final regulations.

STEPS TO DEVELOPING RELATIVE WORK VALUES

The development of the RBRVS has involved a five-step process: (1) obtaining raw survey data on physician work separately for each "major" specialty, (2) fitting data from each specialty onto a common relative value scale, (3) calculating total work based on estimates of pre- and post-service work, (4) mapping work values for surveyed ser-

vices into current procedural terminology (CPT) codes used for payments, and (5) extrapolating work values from surveyed services to nonsurveyed services. Although the final step has become moot as virtually all services have been surveyed, the remaining four merit discussion.

Step 1: Obtain Specialty-Specific Work Values

The first step of the RBRVS is to obtain RVWs for different services performed by physicians. From the very beginning, the consensus has been that the best way to do this is to ask physicians; the issue has been how to ask the question. The Harvard study adopted the principle that the basic physician effort was "intra-service" work, or the work involved in the main part of delivery of a service.

Magnitude Estimation. Intra-service work has been directly assessed by "magnitude estimation," where physician-respondents rate the work of a service as a multiple or fraction of the work involved in a standard service. For example, surgeons compared procedures to the work required to perform an uncomplicated indirect inguinal hernia repair, which was defined as having a work value of 100. If a particular procedure required half the work of the standard, it was rated 50; if it required three times as much work, it was rated 300.

Considering all the alternatives, the claims for the validity of magnitude estimation appear convincing. The RVWs derived from magnitude estimation all show the independent influences of four separate components of work—time, cognitive effort, physical effort, and stress, in ways that understandably differ across different medical specialties.

Specialty-Specific Telephone Surveys. In both Phases I and II, survey data were obtained from physicians through specialty-specific telephone surveys of nationally representative samples. In Phase II, a side-study explored other possible ways of obtaining data. It concluded that introducing interaction among panelists by employing groups led to results diverging from the so-called "gold standard" of the telephone survey results. Nonetheless, Phase III was supposed to use a small-group process. Section 3 of this report examines the relative merits of alternative survey methods for obtaining RVW estimates and the costs of different survey techniques.

Rating Vignettes. Each specialty was asked to provide estimates of intra-service work for about 23 services, presented as "vignettes." Each vignette was a brief description of a patient and the service provided. A number of concerns have been raised about the use of these

vignettes, including who provided the ratings, individual differences in perception of the work involved in performing the vignettes, and the representativeness of the vignettes as exemplars of CPT codes. Each concern has been addressed in criticisms of the Harvard study and in research done by Abt Associates, the Physician Payment Review Commission, and others. The result has been minor modifications to the process in later phases of the Harvard study.

Step 2: Fit Work Values into a Common Scale

Work values have been obtained for different medical specialties, at different times, and using different "standard" services as the basis for magnitude estimation. Variation in any of these factors means that the obtained values are not commensurate; "linkage" procedures are necessary to fit measurements to a common scale. This calibration has been done by declaring services from different specialties as having the same amount of work and employing least-squares regression procedures to produce a set of adjustments to transform specialty-specific survey values to a common scale. We discuss this linkage process in detail below and in Section 4.

Step 3: Calculate Total Work Using Estimates of Pre- and Post-Service Work

The magnitude estimates of intra-service work are converted to a total amount of work by adding pre-service and post-service effort. In Phase I, this was done by first obtaining estimates of pre- and post-service work for a sample of vignettes and then using regression analyses to extrapolate to the other surveyed services.

The establishment of total work from intra-service work was subject to much criticism, largely centering on the definition of what constituted pre- and post-service work and how the various components of this work should be structured for separate assessment. In response to these criticisms, Harvard's Phase II developed refined estimates of pre- and post-service work by defining the work involved in those time periods more precisely.

Step 4: Map Work Values for Translating Vignettes into CPT Codes

After linkage and determination of total work values for all the vignettes, the work values must be assigned to billing codes. Where there was a one-to-one mapping between a single vignette and a sin-

gle code, this was a straightforward task. However, the assignment was not always straightforward because: (1) the translation from vignette to the appropriate code was not subject to an unambiguous set of rules, (2) some vignettes from the same specialty were assigned to the same code, and (3) some vignettes from different specialties were assigned to the same code.

Ambiguous cases were decided by various panels convened for the purpose, so that, eventually, most vignettes were assigned to one code. For services from a common specialty that shared a code, the RVW was calculated as the arithmetic mean of the vignette work values. For services from different specialties that shared a code, the common RVW was calculated as the volume-weighted average (using Medicare Part B data) of work values from the realigned specialty-specific scales.

One type of billing code common to virtually all specialties is evaluation and management (EM) (originally numbered 90000 through 90699 in the CPT coding system). As the RBRVS process evolved, it became clear that the EM codes were not adequate. These codes were replaced, beginning in 1992, by a new set of codes (numbered 99200 through 99499), whose work values were estimated by a panel of Medicare Carrier Medical Directors convened for that purpose. The panel used values for the original EM codes as a starting point and translated work values to the new codes when they believed it appropriate.

OBTAINING MAGNITUDE ESTIMATES FROM PHYSICIANS

To obtain magnitude estimation data for RVWs, one needs to consider who should provide estimates of work values, alternative methods for surveying physicians, and the cost of obtaining data.

Who Should Estimate the Magnitude of Work?

In deciding who will estimate work, the data collector must ensure that respondents are qualified to respond and that the data-collection process is not subject to "gaming" or other biases. The Harvard study randomly sampled physicians from the AMA Masterfile, a process that was questioned by some, although not on the basis of any hard evidence. Several critics have argued that respondents should be drawn by specialty societies, but that course of action runs a risk of conflict of interest. A resolution of the different positions can be found by employing a Universal Provider Identification (UPIN) database about to begin at HCFA. This new database can be used as

the source for physicians performing more than a specified minimum number of services within the set to be surveyed.

Group-Based Methods for Obtaining Work Values

As part of Phase II, the Harvard study investigated the possibility of using small group process methods for estimating work values instead of the telephone survey approach used in Phase I. We suggest, contrary to the conclusions of the Phase II final report, that the data argue for the future use of some small group process, either Delphi or face-to-face, for generating physicians' work values.

Our opinion is buttressed by our examination of the recent social psychological literature of empirical individual and group-based judgment and decisionmaking. We investigated whether tasks similar to assessing relative work values were better suited to collective individual or group methods. Our search yielded 23 published studies, looking at both intellective (where a correct answer can be determined) and judgmental (where there is no a priori correct answer) decision tasks. We view estimating physician work as midway on the intellective-judgmental continuum. The studies were consistent in showing that, strongly for intellective tasks and moderately for judgmental tasks, small-group processing produces greater accuracy and more hypothesis evaluation than individual processing. For both types of tasks, a group advantage occurs without a significant degradation of output caused by differences in member ability and status. This result suggests that future estimates of RVWs should be obtained by some form of group method that permits interaction and feedback among respondents.

The Costs of Different Methods of Data Collection

To better decide which method of obtaining work value data to choose, we developed cost estimates for collecting such data for four data-collection methods: (1) interviewer-administered, one-round telephone survey; (2) one-round mail survey; (3) two-round mail survey (Delphi); and (4) one-round mail survey with a group discussion follow-up.

The cost estimates were for a hypothetical revision of the RVW that would require obtaining assessments for 600 services—50 services each from 12 different panels for the first three methods and 200 services from each of three panels for the discussion method. Sample sizes for the methods were chosen to produce approximately equal between-physician standard deviations of intra-service work values.

The estimated costs showed that the telephone survey was the most expensive method, at $105,000. The single-round mail survey was the least expensive, costing $65,500. The two group methods were approximately equal in cost, with the discussion method ($88,000) costing 10 percent more than the mail survey plus panel ($80,000). The smaller sample size of the discussion method was offset by the travel costs to convene the groups. Based on cost, there is no reason to choose one group method over the other, whereas if an individual-ratings method is chosen, the mail survey has clear advantages over the telephone survey.

LINKAGE

In Phase II of the Harvard study, 275 links were used to align the specialty surveys from both Phase I and Phase II onto a common scale. In producing the common scale, the source individual specialty-specific scales are not rescaled internally, so that the work value relationships within each scale stay the same after adjustment. For the present project, we attempted to replicate the Harvard linkage procedure and developed an alternative linkage procedure based on a different set of assumptions than that used by Harvard.

Replicating the Harvard Linkage Procedure

We attempted to replicate the Harvard linkage procedure. Because only the means and standard deviations of services over physicians were available to us, we could not use, much less validate, the estimation-maximization averaging of the physician-level data used in the Harvard study. In this replication process, we found some problems:

- One specialty, ophthalmology in Phase I, did not have estimated standard errors; we instead used the average standard error for all ophthalmological services in Phase II as a surrogate.
- The Harvard study variance estimates were multiplied by the number of physicians surveyed for each service.
- It was unclear which values were used to center services linked by total work instead of intra-service work.
- The biweight procedure acted as a filter to eliminate some linkages from the regression estimation. Although this was statistically appropriate, it may have implications that have not yet been investigated.

Our replication attempt was largely successful. In general, our results are within 10 percent of the Harvard results except for the three specialties dermatology (Phase II survey), ophthalmology (Phase I survey), and orthopedic surgery (Phase II survey).

A New Look at Linkage

Our experience with the Harvard linkage procedures led us to consider an alternative to their methodology. This alternative, which we term the *perturbation minimization* procedure, is based on four considerations that differ somewhat from those employed in the Harvard study:

- If the links are among services with equivalent amounts of work, then the RBRVS scale should reflect this equivalence.
- Links should be transitive.
- Because a CPT code represents the same work across specialties on average, it constitutes an implicit link.
- Adjusting work values through linkage should preserve as much as possible the originally surveyed work relationships among services.

The perturbation minimization procedure takes place in two discrete steps: redefinition of links and readjustment of scale values after linkage.

Redefinition of Links. We began with our link set equal to the Harvard link set. We then changed this link set in three ways. First, we dropped the Harvard intensity links because these links, which implicitly defined work as the product of time and intensity, seem contradictory. This resulted in the loss of 32 of the 275 original Harvard links that were intensity links. Second, we expanded our link set so that all services with the same CPT code were linked. For any specialty in which a CPT code was surveyed more than once, we formed a new "service" whose work value was the average of the work values of the services with that CPT code. This averaging produced 83 new services. We then linked all common CPT codes across specialties using these averaged services when applicable so that a specialty had no links within its own scale. Common CPT code links were not formed for EM CPT codes 90000 through 90699 because of inherent problems with these codes. Third, we expanded linkages to create transitive subsets of interlinked services, which we termed *orbits*. We created 208 orbits, resulting in a new total of 638 links.

The results of recalculating the least-squares linkage procedures with the expanded link set produced major differences from the original Harvard set. Given that our set of assumptions is as reasonable as that used by Harvard, these results indicate that the linkage procedure is quite sensitive to underlying assumptions, and its validity is consequently questionable.

Readjusting Scale Values. Our proposed revision of linkage requires that all linked services within an orbit have the same work value on the common scale. At the same time, we would like the optimization to ensure that the distances between services within a specialty stay as close as possible to the original surveyed distances. In essence, after redefining the linked service values, we seek to adjust the original values to preserve as much as possible their relationships to the linked and unlinked services in their specialty. We developed a least-squares procedure to do this.

In general, though the percentage changes are smaller than those resulting from the changes in definition of linkage, there are still major differences between Harvard's and the present results. One-fifth of all CPT codes had RVWs that differed from the Harvard values by 15 percent or more. As before, we do not claim that our results should replace the earlier ones, but only that the linkage process is sensitive to methods and therefore should be approached with great caution.

RECOMMENDATIONS

Our objective in this report is not to tell HCFA how the Harvard study should have been done; nor is it to relate what was done correctly and incorrectly in the Harvard study. Instead, our goal is to specify how HCFA best can modify the RBRVS. Such modifications include both short-term fixes to take care of egregious errors, known biases, and omissions from the original set of RVUs and long-term fixes to ensure that the MFS will be a living policy, sensitive to changes in medical technology and the economics of health care.

We recommend that any future magnitude estimation of work values be done using a method that permits individual raters to interact with each other as they estimate RVWs. The two leading candidates for such a method are (1) a two-round mail survey with feedback on the distribution of responses between rounds (i.e., a Delphi process) and (2) a discussion panel preceded by a preliminary mail round. The two methods appear equally valid and do not differ greatly in cost. The benefits obtained in representativeness from a larger respondent sample favor use of the mail survey for long-term changes to the

RBRVS. But for short-term demands, the efficiencies in time and effort from assembling a discussion panel make it the better option.

We recommend that the new HCFA universal provider file be used to select physicians with the necessary experience in the target services for surveys or panels. Representatives from specialty societies can observe and advise panels.

Our examination of the Harvard study's linkage procedures has unearthed a number of technical problems and conceptual ambiguities. We devised an alternative linkage procedure based on a modified set of assumptions and showed that this procedure led to major differences in work values. The conclusion from our analyses is not that our alternative is better than the original linkage procedure, but rather that the links—and therefore any work values that are touched by the links—are very sensitive to changes in the assumptions underlying the procedure. If linkage procedures are to be used in the future, a good deal of further research is required to ensure their validity.

However, we do not believe that linkage need be used for revising the RBRVS. Instead, if a core reference set of RVWs whose validity is not questioned can be found, this reference set can be used to define a common scale for any future estimations of work values, thereby obviating the need to transform an idiosyncratically scaled set of values to a common scale. Because of the importance of any such reference set, its validity must be firmly established by empirically replicable means.

ACKNOWLEDGMENTS

We wish to thank RAND colleagues Margaret Bitzinger for conducting a library search of comparisons group vs. collective individual task performance, Grace Carter for statistical and managerial support, and Gwen Parker for helping provide cost estimates for different measurement techniques. Constructive reviews of an earlier draft by David Kanouse, Daniel Relles, and Sally Trude greatly improved the clarity and accuracy of the report. We thank Edmund Becker, William Hsiao, and Lucian Leape of the Harvard School of Public Health for their very helpful conversations. Acknowledgment is also due to the HCFA Office of Programs and Demonstrations staff, notably Michael Borowitz, Stephen Jencks, and Jesse Levy, for their help and support.

CONTENTS

FIGURE

TABLES

1. INTRODUCTION

BACKGROUND

The Medicare Fee Schedule (MFS), which took effect as mandated by Congress in January 1992, replaces the system of customary, prevailing, and reasonable charges used by the Health Care Financing Administration (HCFA) for physician payment with a system of payment based on relative value units (RVUs). The RVU for a physician's service comprises three elements: (1) the relative value for physician work (RVW); (2) practice expenses, i.e., overhead, excluding malpractice expenses; and (3) malpractice expenses. Under the MFS, payment for a service involves adjusting each of these three elements of the RVU by a separate geographical factor, then summing to produce a value for the service. This RVU is then multiplied by a national conversion factor to yield a dollar amount for payment.

The MFS has potential major consequences for both the total amount and the distribution of Medicare payments for physician services. Because of the large magnitude of expected payment redistributions and the political significance of reforming a major public program such as Medicare, the proposed MFS has attracted and will continue to attract close scrutiny and criticism.

The major innovation of the MFS—and the focus of this report—is the measurement of RVWs. Estimates of physician work under the new fee schedule are based on the Resource-Based Relative Value Scale (RBRVS), which is the product of a major research effort conducted since 1986 by a team of researchers at the Harvard School of Public Health, headed by William C. Hsiao, Ph.D. The "Harvard study" has been conducted in three separate phases through a series of cooperative agreements funded by the Health Care Financing Administration (HCFA). The Physician Payment Review Commission (PPRC) has also played a central role in reviewing the results of the Harvard study as it has evolved, as well as in making independent recommendations concerning improvements and refinements to the MFS. Moreover, as the RBRVS has evolved, it has been extensively commented upon and criticized, and numerous revisions have been recommended. However, the estimates of physician work contained in the MFS rules[1] issued to take effect in January 1992 are based pri-

[1]Health Care Financing Administration (1991b).

marily on the findings of Phase II and part of Phase III of the Harvard study.[2]

OUTLINE OF THIS REPORT

This project examines selected issues regarding the measurement and calculation of RVWs that resulted from the RBRVS. We have examined alternative methods for: (1) obtaining survey data on the amount of work performed for services from physicians in different specialties, and (2) transforming those data into a common scale of relative work values. The work reported here was conducted in 1991, before any release of results of Phase III of the Harvard study or the HCFA final regulations.

Our objective is not to tell HCFA how the Harvard study should have been done; nor is it to relate what was done correctly and incorrectly in the Harvard study. The political reality is that the initial RVWs provided by the Harvard study are—like it or not—here to stay. The next step is to specify, learning from the lessons of the Harvard study and associated efforts, how HCFA best can modify the RBRVS. Such modifications include both short-term fixes to take care of egregious errors, known biases, and omissions from the original set of RVUs and long-term fixes to ensure that the MFS will be a living policy, sensitive to changes in medical technology and the economics of health care.

Section 2 of this report presents a detailed discussion of the development of the RBRVS. It examines the steps employed in creating the RBRVS, identifies limitations in the methods and assumptions related to each step, and notes the major criticisms of the process that have been published.

The next two sections examine specific issues in constructing the RBRVS and present alternative methods that could be used in future revisions of the MFS. Section 3 compares individual- and group-based survey methods for obtaining the magnitude estimation data

[2]In this report, we refer to the complete body of work related to the development of the RBRVS as the "Harvard study," specifying the phase when appropriate. Phase I of the study was summarized in a series of articles in the October 28, 1988, issue of the *Journal of the American Medical Association* (Becker, Dunn, and Hsiao, 1988; Braun et al., 1988a; Braun et al., 1988b; Dunn et al., 1988; Hsiao et al., 1988a; Hsiao et al., 1988b; Hsiao et al., 1988c; Hsiao et al., 1988d; Kelly et al., 1988). The final report of Phase II of the Harvard study is Hsiao et al. (1990); other journal articles are forthcoming. Results from Phase III were presented to HCFA throughout 1991 and 1992. The final report of this phase was scheduled for release in December 1991 but has not yet been made available as of this writing (April 1992).

that provide RVWs. It examines questions concerning who should provide ratings, the method for obtaining measurements, and the costs of obtaining data using alternative methods.

Section 4 looks at the issue of merging surveys from different specialties and possibly over different time periods into a common scale of measurement. The least possible distortion of the magnitude relationships among services is desirable when transforming from a specialty-specific to a common scale. In this section, we examine the methods used to derive the common scale, propose alternative methods, and compare the results of the different methods.

Finally, Section 5 briefly recapitulates the findings of the previous two sections and recommends how to revise RVWs for both the short-term and long-term.

2. DEVELOPMENT OF THE RESOURCE-BASED RELATIVE VALUE SCALE

OVERVIEW

This section describes how estimates of RVWs were developed for use in the MFS and examines several important methodological issues and assumptions related to the process for developing these estimates.[1] Because the results of the last two phases of the Harvard study have not been widely disseminated and evaluated, one important objective of this section is to examine critically the methods and assumptions employed to produce the final RBRVS that is effective for calendar year 1992.[2]

The development of the RBRVS involved a five-step process: (1) obtaining raw survey data on physician work separately for each "major" specialty, (2) fitting data from each specialty onto a common relative value scale, (3) calculating total work based on estimates of pre- and post-service work, (4) mapping work values for surveyed services into codes used for payments, and (5) extrapolating work values from surveyed services to nonsurveyed services. Here, we will look separately at each step.

STEP 1: OBTAIN SPECIALTY-SPECIFIC WORK VALUES

The first step of the RBRVS was to obtain RVWs for different services performed by physicians. From the very beginning, the consensus has been that the best way to do this is to ask physicians; the issue has been how to ask the question. The Harvard study adopted the principle that the basic piece of physician effort was "intra-service" work, or the work involved in the main part of delivery of a service. Although intra-service work as a term of art has evolved slightly over time, basically it means:

- For office-based evaluation and management (EM) services: the face-to-face encounter time;
- For hospital visits: the time spent on the floor;

[1]A detailed table summarizing the tasks and research methods of all three phases of the Harvard study, as well as other studies that have had an influence on the development of the RBRVS, is presented in Appendix A.

[2]Health Care Financing Administration (1991b).

- For surgical procedures: the skin-to-skin time; and
- For laboratory and imaging services: the entire task.

Magnitude Estimation

The Harvard study employed "magnitude estimation" as the primary methodology for obtaining physician work values. Magnitude estimation is a well-established psychometric technique[3] that has been successfully employed to assess subjective values in many different domains. Respondents are given a reference value or values that define a unit of measurement; for the Harvard study, this was the so-called "standard" service (e.g., for general surgeons, an uncomplicated indirect inguinal hernia repair on a 45 year old male), which was defined as having a work value of 100. Then, the respondents rate every other service in the survey relative to the standard or reference value. For example, if a particular service requires half the work of the standard, it should be rated 50; if it requires three times the work of the standard, it should be rated 300.[4]

Direct magnitude estimation is not the only possible way to obtain RVWs. One could instead decide that all work was of equal effort and simply base work on the time spent.[5] Alternatively, one could decompose work into separate parts, somehow estimate the parts separately, and estimate the correct recomposition function to calculate the RVW as a whole. Or, one could employ some form of multiple comparisons technique to first rank-order and then place on a cardinal scale the set of services under scrutiny.[6]

But, considering all of these alternatives, the analyses of Phase I of the Harvard study[7] concerning the validity of magnitude estimation appear convincing. The RVWs derived from magnitude estimation reflect a linear combination of four separate components of work—time, cognitive effort, physical effort, and stress. Furthermore, the weights attributable to these components of work vary across different medical specialties. This result is entirely reasonable, because different specialties are not likely to have the same "mix" of work components. This finding also means that direct magnitude estima-

---

[3]See, for example, Stevens (1957, 1966); Stevens and Galanter (1957).

[4]In a technical sense, this method should be called ratio estimation rather than magnitude estimation (Kahan, 1968).

[5]For example, Maloney (1991).

[6]For example, Bock and Jones (1968).

[7]Hsiao et al. (1988d).

tion of work is a more efficient way to obtain values than measuring the separate components of work and their appropriate mix for each specialty.

Despite some criticism of magnitude estimation,[8] we found no inherent flaw or limitation in the use of magnitude estimation to obtain subjective judgments from physicians about RVWs, and we concur with the PPRC and other recent evaluations[9] supporting the appropriateness of magnitude estimation for this purpose.

We note a major inconsistency in the use of magnitude estimation. One argument for accepting magnitude estimation as a method of estimating work is that the problem of specifying the way the components of work combine need not be addressed. Given this (in our view) strong argument for magnitude estimation, we find it perplexing that the Harvard study chooses, when convenient, to assume that work is the product of time and "intensity," where intensity is implicitly defined as everything else involved in work except time. No evidence was presented in the Harvard study or by anybody else to justify the hypothesis that work equals time multiplied by intensity; this lack of evidence calls into question any analyses that assume this hypothesis to be correct. This issue will appear a number of times in the development of the RBRVS.

Specialty-Specific Telephone Surveys

In both Phases I and II, survey data were obtained from physicians through specialty-specific telephone surveys (18 specialties in Phase I; 15 in Phase II). A nationally representative sample of about 185 physicians was identified in each specialty and contacted; approximately 100 physicians per specialty participated in the surveys. Furthermore, as part of Phase II, seven specialties included in Phase I were resurveyed either because they constituted a substantial portion of services paid for under Part B of Medicare or because of the need for a broader representation of subspecialties or services. In total, the telephone surveys produced 40 separate surveys that were used as input data for developing the RBRVS.

Phase III was supposed to use a so-called small-group process (some combination of mail surveys and face-to-face meetings) instead of

---

[8]See, for example, Pasnak (n.d.).

[9]Physician Payment Review Commission (1991), p. 23.

telephone surveys to obtain work estimates. Following the first two phases, the RVWs obtained through telephone surveys are referred to by the Harvard study group as "gold standards," although no formal evaluation of alternative methods has been conducted by Harvard or others to support this assumption. Section 3 of this report examines the relative merits of alternative survey methods for obtaining RVW estimates, including the costs of different survey techniques.

A statistical algorithm was used to replace nonresponse missing values and values excluded as outliers. For particular services, surveyed physicians might not provide an estimate of work, for example if they did not feel qualified to respond. To estimate a service's mean work value and associated standard deviation across physicians given these missing values, the estimation-maximization algorithm[10] provides a better estimate of the mean and standard deviation than ignoring the missing values or replacing them via some ad hoc procedure.

The estimation-maximization algorithm consists of an estimation step and a maximization step which are alternated until convergence. To begin with, initial estimates for the parameters are calculated based on the existing data. In the estimation step, the conditional expectation of the missing data's contribution to the likelihood function is calculated given the present estimated parameters. In the maximization step, the maximum likelihood estimates of the parameters are calculated just as is usually done when no data are missing. These two steps are iterated until the parameter estimates converge. Although the use of such an algorithm seems appropriate, we were unable to evaluate its effect, and how it was employed, on final RVWs because we did not have the necessary individual physician-level data.

Physician Services Defined Using "Vignettes"

Each specialty was asked to provide estimates of intra-service work for about 23 services, presented as "vignettes." Each vignette was a brief description of a patient and the service provided. Although vignettes were defined independently of codes used for billing purposes, the ultimate goal was to match them to HCFA Common Procedure Coding System (HCPCS) billing codes.[11]

---

[10]See, for example, Dempster, Laird, and Rubin (1977); Little and Rubin (1987).

[11]The HCPCS is used for payment of physician services under Part B of Medicare. HCPCS is primarily based on the American Medical Association's (1991) Current Procedural Terminology (CPT) coding system, with some additional codes defined by HCFA and its fiscal intermediaries.

Criticisms of Data-Collection Procedures

A number of concerns have been raised about the use of these vignettes in Phase I of the Harvard study.[12] Phases II and III of the study were designed to respond to these criticisms and achieved varying degrees of success. Here, we discuss several of the most salient of these criticisms.

Who Performed the Ratings. The Phase I surveys included all physician responses, regardless of the physician's experience in performing surveyed services (i.e., "fitness to rate"). This was a concern because physicians who perform a service infrequently might produce biased estimates of work.

This concern was directly addressed in a study conducted by Abt Associates for the Society of Thoracic Surgeons.[13] This study, which examined services performed by thoracic surgeons, differed from the Harvard study in that the overall specialty was divided into three subspecialties, each of which was independently surveyed. Respondents thus were required to have personal experience with the services they rated. The Abt study concluded that their ratings had greater validity than the Phase I Harvard study ratings of the same specialty.

The Harvard group addressed this potential bias in Phase II by conducting regression analyses using physician characteristics, including frequency of performing a service, to predict physician-specific deviations from the median work rating for each service. The results of this analysis indicated that the physician's frequency of performing a service had no significant effect in explaining these deviations. PPRC reports, however, that its own analyses led to the conclusion that certain services would have had substantially different work values if the responses of physicians who performed the service infrequently were excluded from the surveys.

Individual Differences in Work Perception. A different issue regarding the individuals providing the ratings is that the perceptions of individual physicians might differ, either randomly or systematically, regarding the work involved in the standard service. That is, different people might have different ideas of how much work is involved in "100 units of work." The Phase I surveys assumed that all

[12]The majority of these criticisms are summarized in Physician Payment Review Commission (1989, pp. 37–39) and Physician Payment Review Commission (1991, pp. 24–26).

[13]Noether et al. (1990).

physicians within a specialty viewed the standard service as the same, absolute level of work.

One major refinement in Phase II is that physician estimates of relative work were adjusted to account for different "perceptions" of the standard service. This adjustment process required, instead of the assumption of a common perception of work for the standard service, the assumption that every physician share a common mean work value across the surveyed vignettes. This latter assumption also permitted standard errors to be estimated for the standard service, and the Phase II final report stated that these standard errors "provide a more valid estimate of the standard deviation for each service, including the standard service."

To regard the resulting difference between Phase I and Phase II as an improvement assumes that physicians shared a common average value of work across the 22 to 25 services surveyed. We have no a priori reason to believe that this assumption is valid. Moreover, the assumption of a common assessment of standard services is not really necessary for magnitude estimation, which is concerned with the differences between the standard and measured services.

What remains is to better address the problem of anchoring standard services so that we may safely assume that different raters use a common scale. To date, neither the Harvard study nor any other published comments have addressed this issue.

Vignettes. A number of criticisms of Phase I stated that the vignettes did not reflect the work physicians do. For example, a concern was raised that the vignettes might not measure the work involved in treating Medicare (i.e., largely elderly) patients. This issue was addressed directly, but in a limited fashion, in Phase II. The Phase II report presents evidence for six pairs of vignettes to suggest that intra-service work and time do not vary substantially for Medicare and non-Medicare patients receiving the same service. In addition, more vignettes in Phase II were defined using age 65 and above.

Also, the Abt study challenged the validity of the standard service used for the measurement of thoracic surgery. It also questioned whether the vignettes employed for that survey adequately reflected the nature and variety of work performed by thoracic surgeons. Again, differences between the Abt study and the Harvard study were attributed by Abt in part to refinement of the standard and other vignettes, and the Abt results were labeled more valid than the original ones. In response to Abt, the Harvard group pointed out that most of

the differences between the two studies were related to the pre-service and post-service work estimations, not to the core, intra-service work estimations.

STEP 2: LINK SPECIALTY-SPECIFIC WORK VALUES ON A COMMON SCALE

After assessments of intra-service work had been separately estimated for different specialties, a procedure was needed to compare these assessments on a common scale. We might liken the problem to one of different sets of length measurements using rulers marked with different units of measurement. Some rulers measure in inches, others in centimeters, and still others in feet, yards, or Ångstrom units. To make these measurements comparable, the values on each ruler must be transformed by a multiplicative constant. To find the set of multiplicative constants requires finding objects measured on the different rulers that are known to have the same length.

"Same" and "Equivalent" Links

Two services from different specialties having the same amount of intra-service work—and thus usable for comparing specialty-specific measurements—were called "linked" in the Harvard study. In Phase I, a multi-specialty panel of 24 physicians identified "same" (i.e., involving identical work) and "equivalent" (i.e., involving similar amounts of work) services to serve as links. The process for identifying links was an iterative one, involving both clinical judgment and empirical evidence. The panel originally identified 159 pairs of services as potential linkages but reduced the number to 75 after eliminating pairs whose elements came from nonsurveyed specialties and pairs whose elements differed by more than 25 percent in average time. The final number was increased to 82 (40 same, 42 equivalent) after a cluster analysis on time and work identified additional potential links, seven of which were approved by the multi-specialty panel.

Additional links were developed in Phase II for specialties not included in Phase I, as well as for five specialties from Phase I. These links were developed by multi-specialty panels drawn from 26 specialties. The concept of "same" and "equivalent" links from Phase I based on intra-service work was expanded to include four types of links, based on (1) intra-service work, (2) total work, (3) intensity of work, or (4) intra-to-total work. In five multi-specialty panel meetings, panelists identified 193 pairs of services from 362 potential links.

The total number of linkages, therefore, was 275 (82 from Phase I, 193 from Phase II).

The inclusion of links based on intensity has been criticized by PPRC and others because it assumes that work is a simple product of time and intensity.[14] This linear relationship between total work and time is questionable. Furthermore, it is not clear from the Phase II final report how intensity links were entered into the regression analysis. Finally, the sensitivity of the final common work scale to changes in links, or changes in the work values for links, has not been evaluated.[15]

Regression Model for Linking Services

For both Phase I and Phase II, two specialties were typically connected by more than one pair of linked services, and the multiplicative transformations demanded by each pair were typically not the same. This was deliberately done because of the assumption that there was error of measurement in both the estimation of magnitude and the assumption of equivalency. All of the linked services were combined in a regression analysis to produce an estimated set of transformations to move the specialty-specific measurements to a common scale. This transformation ensured that the original relationships among work values within each specialty survey were preserved.

For Phase II, the regression analysis to produce a common work scale employed input data from 40 specialty surveys representing 33 distinct specialties (15 from Phase II, 3 from Phase I resurveyed in Phase II, 4 resurveyed as special studies in Phase II, and 11 from Phase I). For specialties surveyed in both Phases I and II, results from each phase were treated as separate inputs into the regression.[16]

STEP 3: CALCULATE TOTAL WORK USING ESTIMATES OF PRE- AND POST-SERVICE WORK

Recall that for each service included in the Phase I and II surveys, physicians were asked to estimate their intra-service work and time.

---

[14]See the earlier discussion regarding magnitude estimation.

[15]Each of these issues is addressed in detail in Section 4 of this report.

[16]This regression model and its limitations are discussed, along with a presentation of an alternative approach, in greater detail in Section 4.

However, RVWs are based on the total amount of work. Step 3 provided a means of adding pre-service and post-service effort to the intra-service work to estimate the total amount of work that is the RVW for a service.

In Phase I, estimates of pre- and post-service times were obtained in the original telephone surveys for 55 vignettes. Then, a follow-up telephone survey was conducted among physicians in seven specialties who participated in the original survey to obtain additional estimates of pre- and post-service times. When combined, these surveys produced data on pre- and post-service times for 153 different services. Regression analysis was used to develop estimates of pre- and post-service times for the remaining surveyed services for which only intra-service work and time had been obtained. The estimates of pre- and post-service times were multiplied by estimates of work intensity (i.e., work per minute) to obtain final values of pre- and post-service work. This process produced estimates of total work for all 372 distinct services in Phase I.

Surgical Services

The establishment of total work from intra-service work was subject to much criticism. PPRC[17] conducted a study to separate surgical global service into (1) pre-operation visits, (2) the operation (including scrub work), and (3) post-operation visits. They proposed using intra-service work and scrub work from the Harvard study as the measure of work for the operative component. Estimates of pre-operative and post-operative visits were obtained from specialty societies, who used either a consensus process or a committee process. PPRC regards these values as more valid than the Harvard estimates.

The Abt study subdivided pre-service and post-service work into separate parts and estimated the work value of each part, summing to obtain a total work value. Their results differed in major ways from the Harvard total work values, both in terms of differences in the value of individual services and in terms of a systematic tendency for the Abt values to be of more extreme magnitude (i.e., larger for highly valued services and smaller for less valued services). Abt views the difference between their results and the Harvard results as due to a "compression" artifact of the Harvard method.

In response to these criticisms, Harvard's Phase II developed refined estimates of pre- and post-service work by defining the work involved

[17]Physician Payment Review Commission (1990, 1991).

in those time periods more precisely. Pre- and post-service work was first defined conceptually as eight components, including (1) initial consultation, (2) hospital admission work-up, (3) pre-operative evaluation, (4) other pre-operative work, (5) post-operative follow-up on day of surgery, (6) follow-up visits in intensive care unit after day of surgery, (7) follow-up visits in acute care unit after day of surgery, and (8) post-hospital follow-up visits within 90 days of surgery.

For data-collection purposes, this conceptual model was collapsed into three components: (1) pre-operative, (2) same-day post-operative, and (3) office follow-up. Data on work and time for these components were collected for selected services as part of the Phase II surveys, as well as from specialty panels. A fixed value of 0, 15, or 25 minutes was assigned for other pre-operative work, depending on procedure and setting. The initial consultation and hospital admission work-up were excluded.

In Phase III, the conceptual model of pre- and post-service work was further refined into five components: (1) pre-surgical EM, (2) other pre-surgical work, (3) post-operative follow-up on day of surgery, (4) follow-up visits in hospital after day of surgery, and (5) follow-up visits in office. In this phase, direct estimates of work and time were to be obtained during the pre-operative and post-operative periods for about 300 additional surgical procedures, including the number, duration, and work values of visits before and after surgery. Estimates based on the sum of the five components were to be compared to direct estimates of total work and time for entire global service. Finally, direct estimates of the two major components of post-service work (e.g., before and after hospital discharge) were to be obtained.

Regression Models for Pre- and Post-Service Work

Regression models were developed in Phase II to estimate the three components of pre- and post-service time defined above as a function of (1) intra-service work, (2) intra-service time, (3) hospital median length of stay, and (4) category of surgical service. Six models were used—three for services primarily performed in inpatient settings and three for services primarily performed in outpatient settings.[18] The predicted values of pre- and post-service times obtained from the regression models were then multiplied by the work intensity values for each component to produce a work value for each component of the

[18]The models for pre-service and same-day post-service were physician-level regressions, whereas the models for office follow-up were service-level regressions.

service.[19] The total work value was thus equal to the sum of the work estimates for each component of pre- and post-service work.

The input data used to develop these regressions were not available for evaluation as part of this study. Obviously, these regressions merit examination, especially because Phase II pre- and post-service work values are substantially higher than those obtained by other researchers.[20] Other issues that warrant further study include (1) the assumption of *constant* work intensity across surgical services for each component of pre- and post-service work, and (2) the value of obtaining total work through a "bottom up" approach (i.e., component analysis) compared with direct estimation of total work.

Proposed Studies of Pre-Service and Post-Service Work

Because of uncertainty about the final global fee policy to be adopted by HCFA, services were grouped into three categories: (1) invasive procedures, which include all components of work; (2) endoscopic procedures, which include only work performed on the day of the procedure; and (3) minor procedures, which exclude pre- and post-service work.

PPRC has proposed to validate specialty society data using claims data from carriers that do not include visits in the global fee, data from health maintenance organizations and multi-specialty groups, and physician survey data. Also, PPRC intends to convene a panel of physicians "not directly affected by payment reform" to assess the face validity of estimates for services where the objective data for comparison are inadequate.

EM Services

Data from Phases I and II clearly indicated the inadequacy of CPT codes for EM services (i.e., visits). Physician estimates of work for the same service varied widely across specialties, suggesting that a single, valid, reliable RVW could not be determined for most visit codes.

PPRC has conducted its own study of visit codes and developed detailed recommendations about how visit codes should be modified.[21]

---

[19]The work intensity values for each component were (1) 2.2 for pre-service, (2) 3.0 for same-day post-service, and (3) 2.5 for office follow-up. Other pre-service work was assigned an intensity value of 0.8.

[20]Physician Payment Review Commission (1991), p. 42.

[21]Lasker, Marquis, and Morrow (1991).

The final authority for revising these codes, however, rests with the CPT Editorial Board. The visit codes developed by the CPT Editorial Board reflect some of the PPRC recommendations (e.g., to include encounter time in the visit definition). The final visit codes included in the MFS, however, involve many more distinctions and categories than proposed by PPRC. The structure of the new visit codes is an extremely important area for further research and evaluation but was beyond the scope of this study.

STEP 4: MAP WORK VALUES FOR VIGNETTES INTO HCPCS CODES

After linkage and determination of total work values for all of the vignettes, the work values must be assigned to HCPCS billing codes. Where there was a one-to-one mapping between a single vignette and a single HCPCS code, this was a straightforward task. However, the assignment was not always straightforward because (1) the translation from vignette to the appropriate code was not subject to an unambiguous set of rules, (2) some vignettes from the same specialty were assigned to the same HCPCS code, (3) some vignettes from different specialties were assigned to the same HCPCS code, and (4) most HCPCS codes did not have a vignette assigned to them. The last problem is addressed in Step 5, below; the remainder are addressed here.

Translating Vignettes to HCPCS Codes

A multi-specialty team of technical consultants translated vignettes into HCPCS codes. However, this process was not a straightforward task for the following reasons. Many panelists were not thoroughly familiar with the CPT codes that form the basis of the HCPCS coding system. Therefore, a separate panel of consultants from the AMA staff responsible for the development and maintenance of CPT codes resolved disagreements in assignments by the original panelists. Another translation problem was that some surgical services require multiple codes to be described accurately. In these cases, the work values obtained for the vignette were allocated across more than one CPT code. Finally, codes for EM services were found to have vastly different work values across specialties. Work values for EM services, therefore, were not combined across specialties.

Mapping Vignettes with Common HCPCS Codes

Under the MFS, however, each HCPCS code can have one, and only one, RVW. Therefore, it was important to find a way to combine work values for the same HCPCS code. The Harvard study used different methods for dealing with vignettes mapped to the same CPT codes depending on whether the common codes were found within or across specialties.

For services from a common specialty that shared an HCPCS code, the RVW was calculated as the arithmetic mean of the vignette work values. Although the argument may be made that the underlying assumption that each of the vignette variations on the HCPCS code occur with equal frequency is untenable, no way to estimate the true prevalence of the variations exists. The mean, therefore, seems, *faute de mieux,* to be the right answer. For specialties having many, widely differing, vignettes sharing a common HCPCS code,[22] the solution to any inequity lies in revising the code, not the RVWs.

For services from different specialties that shared an HCPCS code (typically also paired as a cross-specialty link), the common RVW was calculated as the volume-weighted average (using Medicare Part B data) of work values from the realigned specialty-specific scales. For example, if specialty X and specialty Y billed to the same code and specialty X accounted for two-thirds of the billings, then the RVW was two-thirds times the specialty X work value plus one-third times the specialty Y work value. This averaging of work values has the effect of assigning RVWs for vignettes that share CPT codes that are not inconsistent with the original ratios of work within the specialty. That is, continuing our example, specialty X physicians might get less for a linked service than the surveyed respondents believe appropriate because the same service is also provided by another specialty but at a lower estimated work value. In Section 4, we present an alternative definition of linkages and a way to calculate RVWs on a common scale that minimizes this type of departure from the original ratios.

Evaluation and Management Services

One type of billing code common to virtually all specialties is that for EM service (originally numbered 90000 through 90699 in the CPT coding system). As the RBRVS process evolved, inadequacies in the EM codes became apparent.

---

[22]For example, almost all "50-minute hours" in psychiatry share the same HCPCS code of 90844.

A major study of EM services was the PPRC examination of visits and consultations. In this study,[23] 339 physicians in three specialties (internal medicine, rheumatology, and urology) were surveyed to obtain estimates of physician time spent with a patient during the day of the visit for seven categories of EM services (both hospital and office).[24] Physicians were also asked for estimates of the time related to the visit, but performed before or after the day of the visit, for five categories of EM services.[25] Finally, physicians used magnitude estimation to assess the total work for each visit and the proportion of total work performed on the day of the visit. The survey was conducted using encounter forms for 55 consecutive office visits, hospital visits, and consultations. This survey produced encounter forms for a total of 19,143 visits.

PPRC later convened a 46-member consensus panel[26] to examine the following issues: (1) relationship between different measures of time and total work, (2) differences in physician practice styles and use of nonphysician providers, (3) specific EM services provided during visits of a particular duration, (4) impact of specific variables on the relationship between work and time, and (5) common encounter times for different classes of visits. Information was collected from panelists via telephone and mail surveys, as well as face-to-face meetings. The panel used data from Harvard Phase I, the PPRC Survey of Visits and Consultations, and an AMA ad hoc committee on visits and levels of service.

The panel recommended that EM codes recognize three classes of visit (new patient/initial care; established patient/subsequent care; and consultation) and five levels of service within each class based on content and "typical encounter time." PPRC refined the recommendations of the consensus panel into a visit coding system, consisting of 12 classes of visits and five levels of services within each class.[27]

---

[23]Lasker, Marquis, and Morrow (1991).

[24]The categories were (1) record review, (2) history and physical exam, (3) counseling, (4) charting and dictation, (5) contact with other providers, (6) patient-specific contact with house staff, and (7) scheduling activities.

[25]The categories were (1) review of records, (2) talking to patient and family, (3) charting and dictation, (4) contact with other providers, and (5) scheduling and obtaining test results.

[26]The panel consisted of 29 physicians, 5 representatives from the CPT Editorial Panel, 8 representatives from Medicare carriers and private insurers, 2 consumer representatives, 1 nurse, and 1 physician assistant.

[27]The 12 visit classes include (1) new and established patient office visits, (2) initial and subsequent hospital visits, (3) initial and follow-up consultations, (4) initial and subsequent nursing facility visits, (5) initial and subsequent rest home visits, and (6) new and established patient home visits. The levels of service include the following

Final authority for implementing new CPT codes, from which the HCPCS codes are derived, comes from the CPT Editorial Panel of the American Medical Association. The panel proposed new CPT codes in November 1990, and in January 1991 began a pilot test of these new codes for the following EM services: (1) office and outpatient visits, (2) inpatient hospital visits, and (3) consultations. These pilot codes include categories for (1) office and outpatient visits for new patients, (2) office and outpatient visits for established patients, (3) initial inpatient hospital care, (4) subsequent inpatient hospital care, (5) initial consultations, and (6) follow-up consultations. Each category is also divided into 3–5 levels of care. As a result of these efforts, a new set of EM codes were established and assigned the numbers 99000 and upward. RVWs for these codes were established by HCFA in a meeting of small group panels (see discussion of the latest steps at the end of this section).

STEP 5: EXTRAPOLATE WORK VALUES TO NON-SURVEYED SERVICES

Because most services were not directly assessed in Phases I and II, the last step in creating the RBRVS was to develop estimates of total work for non-surveyed services.

Extrapolation Using Charge-Based Ratios Within CPT "Families"

Extrapolation in the first two phases was accomplished by first identifying "benchmark" services from the set of surveyed services that could be identified with a "family" of HCPCS codes. Work values for codes within a family but not surveyed were estimated by extrapolation using the ratio of the average allowed charge for the nonsurveyed service to the allowed charge for the "benchmark" service. In other words, within a family the ratio differences among allowed charges were assumed to represent accurately the ratio differences in work. For EM services, extrapolations were specialty-specific, because the same CPT codes had RVWs that varied widely across specialties.

In Phase I, this process produced RVWs for about 1,400 services, accounting for about 67 percent of total Part B allowed charges and about 80 percent of allowed charges for surgical services. Phase II

measures of encounter time: (1) 10 minutes, (2) 20 minutes, (3) 30 minutes, (4) 45 minutes, and (5) 60 minutes.

refined the definition of CPT families and used more recent data for calculating extrapolation ratios. These steps produced RVWs for 2,024 CPT codes (200 surveyed plus 1,824 extrapolated), accounting for about 84 percent of Part B allowed charges for surgical services. When combined with findings from Phase I, RVWs were calculated for 2,412 CPT codes in 262 families.

In Phase II, extrapolated values were validated by comparing surveyed values with extrapolated values in families with more than one surveyed service. This involved 104 surgical services in 39 families. After excluding extreme values, the average discrepancy between surveyed and extrapolated RVWs was 16.2 percent. Only about one-third of the extrapolated values were within 10 percent of the surveyed values.

Filling the Gaps

Dissatisfaction with the extrapolation method led to Phase III of the Harvard study, which "filled the gaps" in the first two phases by directly estimating work for HCPCS codes that had previously been calculated by extrapolation.

Expert panels of 15 physicians per specialty were established for 26 specialties. Several different surveys of about 50 services each asked for estimates of total work for a standard service and other high-volume services in each family, and intra-service work values for all remaining services in each family, as well as for new services and changing technologies or practice patterns. Although the proposal for conducting Phase III indicated that data would be obtained by some mix of single-round mail surveys and multiple-round group processes, the method actually used has not yet been published.

THE LATEST STEPS

The January 1992 deadline for implementation of the MFS did not allow enough time for assigning RVWs for every HCPCS through Phase III. Also, following the Notice of Proposed Rule Making publishing tentative RVWs,[28] HCFA received tens of thousands of comments recommending revisions to the system. To fill in the final gaps, rectify obvious errors, and establish values for new HCPCS

[28]Health Care Financing Administration (1991a).

codes (including the new EM visit codes), HCFA conducted a series of small-group panels.[29]

Work values for these codes were obtained from panels of HCFA Carrier Medical Directors, using a small-group discussion process and employing estimates of work values from the Harvard study as the starting point for the discussions. Panelists first participated in a mail survey, then in a face-to-face three-day meeting in groups of six members. They were provided with a list of "reference" services, i.e., high-volume services with RVUs that were not under review, and were asked to provide estimates of total work only. The small-group process did not include public voting or consensus on RVWs; the mean value of individual ratings was taken as the group rating. It is worth noting, however, that the RVWs of the individual panel members showed a remarkable convergence to an implied consensus.

[29]These panels were conducted by the HCFA Office of Programs and Demonstrations, based on recommendations of an earlier draft of this report. The results reported here come from personal communications with HCFA staff and observations by the authors.

3. OBTAINING MAGNITUDE ESTIMATES FROM PHYSICIANS

Phases I and II of the Harvard study surveyed, by telephone, a stratified random sample of physicians from the AMA 1986 Physician Masterfile to obtain magnitude estimates of the work required to perform vignettes. For Phase II, several other survey methods were used purely on an experimental basis, including some that involve differing degrees of respondent interaction. The Phase II data used for the published RVWs did not incorporate data from these alternative survey methods. As we noted in Section 2, the Harvard study has drawn criticism for both the source of the sample of physicians and the method of obtaining data from them.

In this section, we examine three aspects of the method for obtaining work values. First, we look at the choice of who should provide estimates of work values. Second, in the main part of this section, we examine different methods for surveying physicians to obtain work values. Finally, we estimate the cost of obtaining data for the major methods examined.

WHO SHOULD ESTIMATE THE MAGNITUDE OF WORK?

Recall that the Harvard study randomly sampled physicians from the AMA Masterfile to obtain survey respondents. This tactic was open to some criticism. First, physicians can declare a specialty on that file without being board certified; some of these self-declared specialists may not have experience with the services being rated. Second, a specialty taken as a whole may be too broad a sampling frame to ensure familiarity with all of the services offered within that specialty. In the study done by Abt Associates for the Society for Thoracic Surgery,[1] that single specialty was divided into three subspecialties, with each subspecialty given its own set of services to rate. Third, a random sample of physicians may not be adequate to the task—instead, peer-recognized experts might be necessary to ensure the knowledge necessary to provide the ratings.

Although these criticisms give rise to thought, no evidence demonstrates that the validity of the Harvard study was compromised by its survey sampling selection. The Abt study provides the strongest case

[1]Noether et al. (1990).

for more select screening of survey respondents; however, that study's ratings of intra-service work did not differ greatly from those of the Harvard study. The large differences between Abt and Harvard for calculations of total work may be due to the different methods of measurement rather than the populations sampled. Admittedly cursory analyses of the effects of physician experience in Phase II of the Harvard study showed no effects.[2] In addition, raters were always free to abstain from responding if they believed themselves unfamiliar with the service portrayed in any vignette.

Several critics have argued that survey respondents and expert panel members who rate services largely provided by a particular specialty should be nominated by that specialty's society, or at least drawn from the appropriate section of the Directory of the American Board of Medical Specialties. The argument is that this is the only way to guarantee that the panel's members will have the direct experience necessary to rate the services. That arrangement is an attractive proposition from the point of view of ease obtaining experts, but it risks conflict of interest. HCFA has a justifiable fear that specialty societies, aware of how the RBRVS process works, could "game" ratings so as to raise their services' work values relative to the values of other societies. Because the MFS is essentially a constant-sum allocation scheme among physician specialties, it is susceptible to such gaming. Anecdotal evidence exists of some gaming attempts during Phase II of the Harvard study. Thus, although specialty societies may well provide the best basis for recommendations to inter-specialty expert panels, as the heterogenous composition of the panels makes gaming more difficult, the societies probably provide a poor source for potential survey respondents for single-specialty panels.[3]

A new database being developed at HCFA may provide the answer to the issue of recruiting future panel members or survey respondents. Beginning October 1991, each physician receiving remuneration from Medicare has a Unique Physician Identification Number (UPIN). Merging the UPIN database with the national historical database that contains all charges to Medicare will yield information about which physicians are billing which CPT codes. This new information can be used to identify experienced physicians who could be tapped as

[2]Hsiao et al. (1990).

[3]Rumors have surfaced that representatives of specialty societies contacted Harvard study panelists before meetings to impress on the panelists the consequences of different types of judgments. Such influence attempts may or may not have occurred, but even their suggestion points to the inherent conflict of interest in having specialty societies play a major role in determining relative work values.

potential survey respondents without reference to membership in the AMA or specialty societies. Suitable geographical, practice setting, and other desirable panel stratification characteristics can be obtained through the Medicare fiscal intermediaries.

The question remains of whether panels should be composed exclusively of procedure performers or of some mix of performers and nonperformers (given that the latter would be familiar with the procedure through education, training, observation, and conversation). We believe that as the RBRVS is revised, some mix of performers and others is necessary. For procedures that are well understood by the general medical community, the heterogenous panel is better for locating relative work on the common, multi-specialty scale of values. However, for procedures that require deeper technical understanding, the performers are uniquely qualified to estimate the RVWs. If these highly specialized services are estimated in the context of a well-defined, previously established scale of values for the more general procedures, the opportunities for gaming are severely constrained.[4]

GROUP-BASED METHODS FOR OBTAINING WORK VALUES

As part of Phase II, the Harvard study investigated the possibility of substituting small groups in place of the telephone surveys for estimating RVWs. This investigation was to assess the validity of the single-round telephone survey method and to explore less costly and more efficient alternatives.[5] The study found that the more the panelists could interact, the further their RVWs deviated from the figures provided in Phase I. The Harvard study concluded that this deviation was a bias resulting from group processes and that an individual-based method (telephone or mail survey) was preferable to other data-collection methods. We agree in general with this test of methods. Upon examination of the results,[6] however, we question whether the Phase I telephone survey results should be taken as a gold standard for method comparison. Absent any evidence that such a gold standard was established, we examine the social psychological literature on judgment of numerical estimations to see how it might guide method selection. That literature argues for the use of small groups, either Delphi or face-to-face, for revising estimates of RVWs. This

[4]Even for sessions rating highly technical services, one or two panel nonspecialist panel members who participate to only a minor extent might assure that the specialists' ratings remain within the bounds of reason.

[5]Hsiao et al. (1990), p. 669.

[6]Hsiao et al. (1990), Chapter 11.

conclusion contrasts with the Harvard conclusion that a single-round mail survey, producing results closest to the telephone survey "gold standard," will suffice.

In this subsection, we first summarize the comparison of three methods of obtaining RVWs conducted as part of Phase II of the Harvard study. Then, we review the recent social psychological literature on collective-individual versus group-based judgment methods, concentrating on the results of empirical studies. We conclude with a reinterpretation of the Phase II small-group process data and recommend how to collect RVWs in the future.

Individuals Versus Small Groups in the Harvard Study

In Phase II of the Harvard study, three methods for generating physicians' estimates of RVWs were compared with values obtained (primarily in Phase I) from national telephone interview surveys. Three panels of general surgeons were selected from a pool of 60 nominees. Each of the six major regional surgical societies submitted 10 nominees to this pool, resulting in a mix of academic and community-based surgeons. Eleven of these surgeons made up Panel A and participated in a combined Delphi and face-to-face group. Nineteen surgeons formed Panel B and took part in a Delphi group with multiple rounds of ratings interspersed with feedback. The 29 Panel C participants completed a single-round mail survey.[7]

Delphi Process. A Delphi method is a quasi-small-group process where participants receive anonymous and limited feedback on each others' ratings.[8] Typically, participants in a Delphi method are surveyed in the following steps:

1. Each individual panelist responds to the questions. The participants' answers are then summarized by the average and a distribution of responses.
2. The results of the first round are returned to the individual panelists for an iteration. After reviewing the feedback of first-round results and the relative location of their individual initial responses with respect to those of their peers, the panelists may modify their beliefs.
3. Multiple iterations of the feedback process typically produce convergence to a consensus.

---

[7]One nominee, originally slated for Panel B, dropped out of the study.

[8]Dalkey, Brown, and Cochran (1969).

This method has at least three advantages. First, a Delphi method may yield more accurate judgments than the combined results from a single-round survey of individuals.[9] This occurs because respondents can, anonymously and therefore without loss of face, revise their judgments. Second, the multiple-round process may produce more ideas than conventional face-to-face discussion groups. Delphi panelists are not subject to any face-to-face social influence processes that might constrain the development of alternative judgments.[10] Third, Delphi transactions do not require the synchronicity of telephone or face-to-face meetings and therefore are not as expensive to conduct.

The major disadvantage of a pure Delphi process is that members are not able to discuss the reasons for their decisions and revisions among themselves. Tasks that involve a number of facts to be considered or recalled are well suited for freely interacting groups. Members' shared knowledge and error-checking discussions serve to increase the accuracy of these groups over groups with limited communication[11] and over collective-individual results.[12]

In the Harvard study, surgeons in Panel B rated the intra-service work of 55 vignettes in three Delphi rounds by mail. Thirty-eight of these vignettes were taken from the Phase II resurvey of general surgery; the remaining 17 were generated in a similar manner.

Combined Delphi and Face-to-Face Process. Panel A combined features of a Delphi method with a face-to-face discussion leading to individuals' re-rating of vignettes. Panelists performed two rounds of Delphi by mail before meeting together. At the meeting, they were instructed to try to reach a consensus. After one discussion session and ratings ("Round 3"), the panelists met for a second time ("Round 4") to hammer out a consensus on the few vignettes for which consensus had not yet been obtained. Panels A and B rated the same 55 vignettes.

Single-Round Mail Survey. The simplest method investigated in Phase II of the Harvard study was a single-round mail survey; as such, it was not really a group process (although labeled as such in the report) but a survey of individuals. Participants completed a sur-

---

[9]Ibid.

[10]For example, Burton (1987); McGrath (1984).

[11]For example, Laughlin and McGlynn (1986).

[12]For example, Michaelson et al. (1989); Stephenson, Clark, and Wade (1986); Vollrath et al. (1989).

vey and their individual responses were averaged to produce a central statistic that was considered representative of the group.

The obvious advantages are in the simple logistics and lowered costs. A single-round mail survey is less cumbersome and has fewer transactions than telephone interviews, successive mailings, or face-to-face meetings. If data from such a mail survey do not differ from data that would be obtained from a more expensive method, the choice of a mail survey is clearly appropriate.

The surgeons in Panel C did not receive the same booklet as the participants in Panels A and B; instead, they rated 25 vignettes, only some of which were from the set of 55 rated by the other panels. They also rated pre- and post-service work for those vignettes. Consequently, their results could be compared only to the national survey.

Results of the Harvard Comparison. For both Panels A and B, feedback about the prior collective-individual ratings tended to produce, as anticipated, a convergence to a consensus. Outliers tended to move toward the center on successive rounds and the differences between the ratings given by the individual surgeons were dramatically reduced. However, during the interactive process, the convergence point diverged over iterations from the telephone survey baseline average. That is, the absolute "degree of disagreement" between a given round and the national survey over all surveyed services increased as the number of rounds increased. Typical disagreement between panel and national survey ratings are demonstrated by the work value differences for the general surgery specialty given in Table 1. For both Panels A and B, the later the round, the more distant the panelists' median judgment from the national survey values.[13] Consequently, the feedback that resulted in the reduction of differences between in-

Table 1

Percentage Absolute Difference Between Phase II Group and National Survey General Surgery RVWs

Panel	Round 1	Round 2	Round 3	Round 4
A	12.6	14.6	21.0	23.8
B	11.0	11.0	15.2	
(A vs. B	12.8)			

[13]Hsiao et al. (1990), Table 11.5, p. 697.

dividual panelists contributed to an increase in differences between the panels and the national survey. Furthermore, the combined method that allowed knowledge-sharing and error-checking discussions generated the greatest differences from the national survey in estimates of work values. In contrast, the collective intra-service work ratings from the Panel C single-round survey compared *favorably* with those of the national survey.

The Harvard study interpreted these differences as a deviation from the national survey "gold standard" and a reason to *reject* multi-round group methods. Our view is that a deviation is not necessarily invalidating; the national survey standards may well be less accurate. Without a true standard, the comparison alone cannot determine which set of figures are the more valid. Unfortunately, there are no other studies comparing individual vs. group decision processes that employ physicians as subjects or address the value of work. Therefore, to shed some light on this issue, we reviewed the recent social psychological literature to examine studies directly comparing individual-based collective decisions to group-based ones.

A Review of Individual Versus Group Judgment Methods

We reviewed the empirical literature of the past ten years for comparisons between individual and group judgment methods. Studies older than ten years of age can generally be found in textbooks and so provide a framework rather than new information, so we focused on specific recent explicit comparisons of the two types of judgment methods.

Intellective Versus Judgmental Tasks. A potentially important distinction in categorizing decision tasks is that made by Laughlin and his coworkers of intellective versus judgmental tasks.[14] Generally, intellective tasks and judgmental tasks are considered to be at opposite ends of an abstract fact-verification continuum. Intellective tasks are those for which a correct answer can be demonstrated, whereas judgmental tasks have no demonstrably correct answer, but instead generate responses based on the beliefs, feelings, or guesses of the decisionmakers.[15] One can illustrate the continuum by looking at the types of tasks that have been investigated:

---

[14]Laughlin (1980); Laughlin and Futoran (1985); Laughlin and McGlynn (1986). See also McGrath (1984); Stasser, Kerr, and Davis (1989).

[15]The distinction between intellective and judgmental is itself not always crisp. For example, although in baseball there are explicitly written rules about what makes a

Intellective pole

Almanac-type questions[16]

Logical rule induction

Learning (e.g., recall) tasks

Jury decisions

Personnel decision tasks

Moral choice dilemmas

Judgmental pole

If the purpose of comparing individual and group decision tasks is the accuracy of the decision, then the criterion for intellective tasks is easy; the right answer is generally known, if not necessarily by the group members. For almanac-type questions, one looks in the almanac and for recall tasks, the list of items to be recalled is known by the instructor or experimenter. However, for judgmental tasks, where there may not be a "correct" answer, the measure of accuracy needs to be defined and may be fairly indirect. Each experimental laboratory has defined its own measure of comparison for studies of judgmental tasks, and they must each be considered separately.

The presence or absence of an empirically verifiable correct answer has led to different predictions about the superiority of individual versus group decision processes. When a correct answer may be stated, then a "truth wins" type of decision rule may be adopted by the group, so that if one member can provide that answer, the group will move to it. Therefore, for intellective tasks, groups should be superior to individuals. But when no verifiably correct answer is known, groups might be more vulnerable to the types of noninformationally based social influence processes hypothesized to occur in conventional face-to-face discussion.[17] For example, some panel members may sway the opinion of other members who previously held better judgments. According to this logic, a survey of individuals should produce better decisions than a panel dealing with the same issue.

---

pitch a ball or a strike (therefore, especially in these days of instant replay, making the task appear to be an intellective one), the reality for the pitcher and batter is that a pitch's status is strictly determined not by its physical location but by what the umpire says it is (making the task a judgmental one).

[16]For example, "What is the height of Mount Kilimanjaro?"

[17]For example, Janis (1972); Stasser, Kerr, and Davis (1989).

An important question is whether the estimation of RVWs is more of an intellective or a judgmental task. Generally, quantitative estimation tasks such as this have been placed within a typology of intellective tasks,[18] but the question merits deeper discussion. On the one hand, because, like jury verdicts, the amount of work to perform a medical service is what a panel declares it to be, the task may be thought of as a judgmental one. On the other hand, there exist empirically observable features of work (such as the time to perform it, the time needed to learn how to perform the service, and the complexity of the procedure) that lead to the existence or more correct versus more incorrect possible work values. In addition, some potential work values may be considered egregiously incorrect, as when the work to perform two separate tasks is less than the work to perform just one of those tasks.[19] These characteristics of the task pull it toward the intellective pole. Our understanding of the task leads us to believe that estimating RVWs falls in the middle of the intellective-judgmental continuum, possibly slightly toward the intellective side. Others might have differing viewpoints on this matter; in any event, it seemed important to examine both intellective and judgmental tasks.

Results of the Literature Search. Our literature search for recent empirical tests of collective-individual versus interacting-group decisions yielded 23 journal articles published within the last ten years. Of these, seven were intellective tasks and 16 were judgmental.

The seven intellective task experiments included three rule induction experiments, one recall of a simulated police interrogation, one recall of a mock trial, one applied problem-solving in a contextually relevant work setting, and one competitive resource-sharing game.[20] In six of these seven studies, the results showed interactive group processing to be superior to the aggregated individual outputs for these intellective tasks. Groups induced a greater number of correct rules than individuals as a result of better hypothesis-evaluation and error-checking. Groups were better than the average individual at recalling information from given scenarios. And group outputs were more accurate than the best or average individual decision when the tasks involved solving contextually relevant and consequential problems related to work. Only one of the seven studies yielded results not favoring group-based decisions over collective-individual

[18]McGrath (1984).

[19]Such instances have arisen in early versions of the RBRVS.

[20]The studies were, respectively, Laughlin and Futoran (1985); Laughlin and McGlynn (1986); Tindale (1989); Stephenson, Clark, and Wade (1986); Vollrath et al. (1989); Michaelson, Watson, and Black (1989); and Irwin et al. (1988).

ones, and the results there were equivocal. In that task, where individuals and groups of two or three persons played mixed-motive games, individual males and female dyads were able to make more correct decisions than male dyads, female individuals, or three-person groups of either gender.

The observations of group process in these studies of intellective tasks further supported the conclusion of the superiority of group-based decisions. In addition to generating greater accuracy and better hypothesis evaluation, group members tended not to be highly confident of incorrect answers. Rather, group members working on intellective tasks generally succeeded at error-checking and convincing other members to select the correct answer, in keeping with the "truth wins" type of social decision process hypothesized for such tasks. Moreover, the intellective task group outcomes were not affected by member ability or status differences when those differences conflicted with a correct answer. For example, in their study of 222 project teams, Michaelson et al. found that not only did all groups outperform their average member, but 215 of the groups outperformed their *best* member with respect to a comparison of group scores and best individual scores.

The 16 judgmental tasks include seven studies of mock jury decisions,[21] five studies of moral-choice dilemmas,[22] one study of election decisions,[23] and three studies of cognitive bias.[24]

Mock jury paradigm. A mock jury study assembles a group to play the role of a jury and presents them with evidence in some abstract form, such as a written summary or a videotape. The jury may be asked for individual estimates of guilt or innocence (for criminal cases) or amount and degree of liability (for civil cases), as well as about confidence in their estimates. The mock jury may then deliberate to a formal verdict or further individual judgments. The group versus individual comparison for this paradigm was the most mixed; further scrutiny showed that the conclusions depended heavily on the particular means of comparison. The mock jury studies that tested possible biasing effects of group process on jury deliberations found little evidence for such effects and concluded that group decisions did

[21]Bankart and Powers (1986); Davis et al. (1984); Davis et al. (1989); Hinsz et al. (1988); Kerr and Huang (1986); Ono and Davis (1988); and Tindale et al. (1990).

[22]Dukerich et al. (1990); Meyers (1989a, 1989b); Nichols and Day (1982); and Turner, Wetherell, and Hogg (1989).

[23]Stasser and Titus (1987).

[24]Glisson (1987); Stasson and Davis (1989); and Stasson et al. (1988).

not differ from what would obtain from taking averages of individual decisions. For example, the mock jury studies by Davis and his coworkers tested the effects of the distribution of individual opinion and the order and timing of the polling sequence on the final verdict outcomes of six-person juries. Individuals' personal opinions were obtained before placing them in groups. Changes in group member opinion were best characterized as occurring at the group level, not as a result of any suspected biasing factors such as exaggeration of individual responses, sequence of individual expression or opinion, or social pressure of majority influence. The studies showing superior results for groups examined the number and quality of arguments presented in the group discussions, which were more and better than those posed by individuals. The single study favoring individual judgments found individuals to be "more fair" than the groups by awarding more similar amounts to male and female victims for the same case; the authors note that the result may be in part due to a requirement that the group decision be unanimous.

Choice dilemmas. A choice dilemma problem is one in which two conflicting values are posed as mutually exclusive choices. For example, in a risk dilemma, security and modest worth may be set against risk and high gain as a person is asked to choose between two different jobs. For another example, in a moral dilemma, individual need may be set against social rules as a person has to choose between obeying the law and helping a relative in distress. For such tasks, where there is no correct answer in any real sense of the term, investigators have looked at the process by which decisions are made. A "good" process is one in which particular individuals do not influence others by virtue of their status or dominance of the conversation, but where influence is through the variety and quality of the discussion. For example, Meyers examined the relative contributions of individuals' previous opinions and the number and nature of discussion arguments in predicting the same individuals' later opinions and found that the variety of arguments explained the group decision more than either individuals' previous opinions or the number of times an argument was expressed.

Election decisions. The study of election decisions considered how well information was shared in a group considering candidates for an election. When most of the information about the candidates was dispersed (privately known) across the group, more information was publicly shared than when most of the information was shared before the group meeting. Moreover, individuals recalled more information supportive of the group decision than contrary to the group decision

in a post-decision recall task. The authors concluded that the face-to-face discussions were a poor way to share information.

Cognitive bias. The studies of cognitive bias examined the relative susceptibility of groups and individuals to factors that bias cognitive judgments. These factors include commitment to the task at hand as well as the various cognitive biases examined and explored over the past 20 years in the cognitive psychology literature.[25] The Glisson study found that the commitment to task of workgroups could not be characterized by the aggregated level of commitment of their individual members but instead was a function of the variety of skills represented by the members of the group. The study by Stasson and coworkers favoring groups showed that groups produced more and better arguments, which influenced cognitive judgments. The study showing no advantage for either individuals or groups showed that groups and individuals were equally susceptible to cognitive biases such as availability, representativeness, and anchoring.

Implications. Table 2 summarizes the results of the literature review. For intellective tasks and to a lesser extent also for judgmental tasks, group-based decision processes were preferred to collective-individual processes. The finding for intellective tasks was as anticipated, but the similar finding for judgmental tasks is mildly surprising. It appears, as one study put it, that instead of groups being subject to feared undue influences, "a heretofore unsuspected robustness against certain procedural influences during group consensus achievement is heartening."[26]

Although none of the studies surveyed is a close match to the task of estimating RVWs, the conclusion suggests that, in the absence of more directly focused evidence, group-based methods (either Delphi or mixed) that permit the raters to interact might be a better choice for revisions to the RBRVS than the mail survey of Panel C or the national telephone survey used for the bulk of Phases I and II. To the extent that producing RVWs is regarded as an intellective task, this could be considered a strong recommendation. If producing RVWs is considered more of a judgmental task, the recommendation is—because of the lack of an empirically defined "correct" answer—weaker but still in the direction of groups. Note that this recommendation is for group interaction before individual ratings, not for group consen-

[25]For example, Kahneman, Slovic, and Tversky (1982).

[26]Davis et al. (1989), p. 1011.

Table 2

Collective-Individual Versus Group-Based Methods

Study	Decision Type	Favored
	Intellective Tasks	
Laughlin and Futoran (1985)	Rule induction	Group
Laughlin and McGlynn (1986)	Rule induction	Group
Tindale (1989)	Rule induction	Group
Michaelson, Watson, and Black (1989)	Recall; problem solving	Group
Stephenson, Clark, and Wade (1986)	Recall; interrogation	Group
Vollrath et al. (1989)	Recall; mock jury	Group
Irwin et al. (1988)	Shared resources	Neither
	Judgmental Tasks	
Bankart and Powers (1986)	Mock jury	Individual
Davis et al. (1984)	Mock jury	Neither
Davis et al. (1989)	Mock jury	Group
Hinsz et al. (1988)	Mock jury	Neither
Kerr and Huang (1986)	Mock jury	Group
Ono and Davis (1988)	Mock jury	Group
Tindale et al. (1990)	Mock jury	Group
Dukerich et al. (1990)	Moral choice	Group
Meyers (1989a)	Moral choice	Group
Meyers (1989b)	Moral choice	Group
Nichols and Day (1982)	Moral choice	Group
Turner, Wetherell, and Hogg (1989)	Moral choice	Group
Stasser and Titus (1987)	Election decisions	Individual
Glisson (1987)	Cognitive bias	Group
Stasson and Davis (1989)	Cognitive bias	Group
Stasson et al. (1988)	Cognitive bias	Neither

sus ratings. The mixed evidence for the mock jury studies, where group judgments replaced individual ones, suggests caution before attempting to insist that the interacting groups reach a consensus.

Whether face-to-face or Delphi methods are employed for revising the RBRVS probably does not matter too much in terms of affecting the values obtained, given that these two types of method produced similar degrees of deviation from the individual-based methods in the Phase II experiment. Other factors such as ease of obtaining the sample of physicians and the cost of obtaining data should be the primary consideration.

THE COSTS OF DIFFERENT METHODS OF DATA COLLECTION

Future revisions of the RBRVS will necessitate obtaining RVWs from physicians. To better decide which method of obtaining RVW data to choose, we developed cost estimates for collecting such data from physicians, using four data-collection methods:

- Method 1: interviewer-administered, one-round telephone survey.
- Method 2: one-round mail survey (self-administered questionnaire).
- Method 3: two-round mail survey (self-administered questionnaires).
- Method 4: one-round mail survey (self-administered questionnaire) with panel follow-up.

We assume that the data collection would be to revise and update the RBRVS and not to restructure the whole scale. Therefore, we assumed that 600 RVWs would have to be obtained. We further assumed that the panels would ask only for the total work for a service and that the RVWs would be provided on the existing common scale of measurement, thereby eliminating the need for links. For each method, we determined a sample size that would produce approximately equivalent between-physician standard deviations of work values.[27] Following a general description of the data-collection methods, we present estimated costs for each method and outline the assumptions for these estimates.

Data-Collection Methods

Method 1: Telephone Survey. This data-collection method follows the "gold standard" used by the Harvard study for Phase I and Phase II of the national surveys. It features a series of one-round telephone surveys administered by trained interviewers, with each survey yielding 100 completed protocols. Before the actual interviews, respondents receive a survey packet made up of a copy of the survey and an endorsement letter. To enhance response rates, respondents receive in advance a check for $20 for their participation in an

[27]Ideally, we would like to have an estimate of between-group standard deviations. But all methods developed so far have only had one group per specialty.

estimated 40 minute telephone interview to rate 50 services.[28] Interviewers administer the survey over the telephone, recording responses on a paper-pencil instrument, which is edited for data entry.[29] Personal computer software and optical scanning equipment is used to log completed surveys and track survey progress.

Method 2: One-Round Mail Survey. This data-collection method involves mailing respondents self-administered questionnaires, to cover the same services as asked for in Method 1. The goal is to have 100 completed forms for each of 12 surveys. To encourage an optimum response rate, respondents receive an initial telephone call, followed in the mail by an initial survey packet made up of the survey, a personalized cover letter and endorsements, and a check for $20. About a week after the initial mailing, respondents receive a short reminder. Four weeks after the initial mailing, nonrespondents receive a replacement survey packet and follow-up telephone calls. Personal computer software and optical scanning equipment are used to log returns and track survey progress. Completed surveys are validated and edited for data entry.

Method 3: Two-Round Mail Survey. This methodology is an adaptation of the model used in Harvard's Phase III survey, which involved a two-round mail survey of small panels in each specialty, with telephone follow-up. Because of the anticipated reduction in variance of response in the second round, a sample size of 50 completed (both rounds) protocols per survey instead of 100 is required. In the first round of the survey, Method 3 is identical to Method 2. Following initial analysis of the Round 1 survey, a second survey is distributed. Respondents to the second round receive an additional incentive payment of $20. Follow-up to the initial mailing of the second-round survey is identical to that described for the single-round survey of Method 2.

Method 4: One-Round Mail Survey with Panel Follow-Up. Method 4 involves the convening of three separate panels of 13 physicians drawn from the specialties of interest. Following the selection and enrollment of panel members, each panel member rates 200 different services. After initial analysis of the data from the mail survey, the panels convene for a one-day meeting to rerate all 200

[28]The Harvard study interviews lasted no more than 40 minutes; asking any more time of a physician will drastically increase the refusal rate.

[29]As an alternative, the instrument could also be set up for computer-assisted interviewing. However, the cost of programming 12 different instruments for a relatively small sample argues for the paper-pencil option.

services and reconcile discrepancies from the first round. To facilitate this process, panelists receive the results of this first round before the panel meeting. Panelists receive an honorarium of $500 for their participation.

Common Features

Several features are common to these data-collection methods:

1. Sample. Samples for Methods 1, 2, and 3 are drawn from the AMA Masterfile from specialty groups or from the UPIN/historical files. Panel members for Method 4 are selected from a similar file by an as yet undetermined basis. The file is assumed to have current addresses and telephone numbers and requires tracking and updating of addresses or telephones at a rate of no more than 5 percent of the sample.

Though sample sizes differ for each method, according to the predicted completion rate or method employed, each method provides ratings of 600 physician services.

2. Completion Rate and Initial Sample Size. Both the size of the initial samples and the expected completion rates vary by method, as indicated in Table 3.

Though the targeted completion rates are somewhat high for surveys of physicians, these rates are achievable given the anticipated high level of motivation and the range of respondent incentives, as described below. By comparison, the Harvard study reported 62 and 72 percent completion rates to the Phase I and Phase II telephone sur-

Table 3

Completion Rates and Sample Sizes

Collection Method	Completion Rate	Initial Sample	Completed Protocols		Completed Items	
			Total	Per Survey	Total	Per Survey
Telephone	0.69	1740	1200	100	600	50
1-Round Mail	0.74	1620	1200	100	600	50
2-Round Mail						
Round 1	0.74	900	667	58	600	50
Round 2	0.90	667	600	50	600	50
Panel	n/a	n/a	39	13	600	200

veys, which did not involve monetary incentives for respondents.[30] Moreover, others have reported rates as high as 78 percent to mailed surveys of physicians that featured advance payment of $20.[31]

3. Survey Instruments. Methods 1, 2, and 3 involve the administration of 12 separate surveys, each of which provides RVWs for each of 50 services. Thus, data for 600 separate services are obtained. In addition to these 50 items, approximately 12 additional items on physician and practice characteristics are collected, for a total of 62 items per survey. At an estimated rate of two items per minute plus general introduction, the estimated time required to complete each survey is 40 minutes, a figure comparable to Harvard study interview times. For Method 4, the 600 services are measured by having three panels each rate 200 services.

For budgeting purposes, we assume that the instrument development task involves selecting measures already developed in Phases I or II rather than composing totally new measures. For all methods, a small pretest (n = 25) has been included in the cost estimates.

4. Respondent Incentives. To ensure high completion rates, a variety of incentives are offered to respondents. All methods provide payment for participation. Though the $20 payment for Methods 1, 2, and 3 does not adequately compensate physicians for their time, even such a modest payment has been shown to have a significant effect on response rates in surveys of physicians.[32] These three methods also feature personalized letters of endorsements from medical associations such as the AMA and specialty societies. The two mail survey methods also include preliminary calls, and (for nonrespondents) extensive mail and telephone follow-up. Method 4 includes many of these features, along with a lump sum honorarium of $500 to each panel member.

Cost Estimates

The estimated costs shown in Table 4 are based on the following general assumptions:

---

[30]Hsiao et al. (1988a); Hsiao et al. (1990).

[31]See Berry and Kanouse (1987).

[32]Berry and Kanouse (1987).

- All estimates are in current (1991) rates and dollars.
- Though overhead is not included, labor costs do include merit and fringe benefits.
- The estimates are only for the data-collection and entry activities (e.g., questionnaire development; hiring, training, and supervising field staff; postage; telephone charges; respondent payments). They do not include the cost of analysis, report writing, or overall project management at the principal investigator level. We assume that the panels in Method 4 are conducted by the principal investigators.
- Travel costs for the panel assume that meetings are in Chicago and that panelists are from different parts of the country. For the one-day meetings, we assume two days/nights of per diem. Travel costs for two investigators are included in the estimates.

Table 4

Summary of Estimated Cost of Data Collection (in 1991 dollars)

Collection Method	Total Cost[a]	No. of Completes	Cost per Complete[b]	Cost per Rated Service[c]
Telephone	$105,000	1200	$87.50	$175.00
1-Round Mail	$65,500	1200	$54.58	$109.17
2-Round Mail	$80,000	1267[d]	$63.14	$133.33
Panel	$88,000	n/a	n/a	$146.67

[a]Total cost of data collection includes all field activities (e.g., interviewing, survey distribution, data reduction), supervision, management, and instrument/materials development.

[b]Cost per complete is derived by dividing the total cost of data collection by the number of completed cases. (This calculation is not applicable to the panel-rating methodology.)

[c]Cost per service is derived by dividing the total cost of data collection by the 600 rated services.

[d]667 completes for the first round and 600 completes for the second round.

CONCLUSION

The results of our examination of how to obtain RVW estimates for revising the RBRVS indicate that methods that permit the raters to interact with each other while rating services are probably preferable to methods that do not. The differences between group-based methods and individually based methods shown in the Harvard study do *not* automatically lead to the conclusion that group-based methods

are flawed, only that the two types of method lead to nonequivalent results. Our survey of the social psychological literature suggests that the group-based methods produce values more indicative of the respondents' true judgments than do the individual-based methods.

The cost estimates of the various data-collection methods show that the panel (Method 4) costs only 10 percent more than the 2-round mail survey (Method 3), a difference we regard as small enough so that relative costs of the two group-based methods need not be a major consideration in choosing between them. We recommend the panel method over the Delphi because it provides more information for the participants.

If an individually based method continues to be the instrument of choice, the Harvard study results show that mail surveys and telephone surveys produce equivalent results. Our cost analysis shows that a single mail survey is about five-eighths of the cost of a telephone survey, a savings that clearly makes it the preferred method.

4. LINKAGE

This section examines in some detail the linkage procedure used by the Harvard study group. We employed our understanding of linkage to attempt to replicate the Phase II results. On the basis of considerations that arose during the replication effort, we constructed an alternative linkage procedure using a "perturbation minimization" concept of linking the diverse specialty surveys to a common scale. We designed this alternative not as a definitive replacement to the Harvard linkage technique but rather as a different method based on slightly dissimilar but equally justifiable assumptions. The perturbation minimization technique involves both a reconsideration of the definition of a link and a modified optimization procedure that can adjust values within each specialty survey. We close with a discussion of the implications of the exploration of this alternative approach.

HARVARD LINKAGE PROCEDURE

The services to be linked across specialties were chosen in a series of steps:[1]

1. Technical consulting groups and the project team developed lists of potential links from the services included in all of the surveys.
2. The cross-specialty panel identified same services (the process, the time, and the type of patient were essentially the same) and equivalent services (intra-service work essentially the same and in the same service category) from the potential links.
3. Potential links that differed by more than 25 percent in terms of intra-service time were discarded.
4. Services were classified into service and setting categories and additional potential links were identified from these clusters. The cross-specialty panel chose further links from this list.

This method yielded 275 links of paired services.

The goal of the linkage procedure is to align the individual specialty scales of work to a common scale so that services from different spe-

[1]Braun et al. (1988b).

cialties with the same values on the common scale have the same level of work. Three key assumptions were made in this alignment:

1. The sampling of services in the various surveys is representative of all the services provided by a specialty, so that the mean ratings of work from a specialty survey approximate the mean rating for work for all services performed by that specialty.
2. When the services in different specialties are judged to be the same or equivalent, they involve nearly the same amounts of work.
3. Each specialty's scale as a whole is unchanged after alignment so that the ratios of all the services within the specialty remain the same.

The first assumption translates into the use of physician-level mean ratings. The second assumption says that the chosen links are reasonable. The third assumption means that the scales will be aligned by shifting them relative to each other. The individual scales are not rescaled internally first nor are individual services within a specialty shifted different amounts. A specialty scale stays rigid as it is shifted relative to the other scales.

Let d'_{ij*} be the adjusted average of the physician-level logarithms of the work ratings, where the average is taken over the approximately 100 physicians surveyed for a specific service i in specialty j. This average was calculated in the Harvard study using the estimation-maximization procedure,[2] as discussed in Section 2.

The adjusted d'_{ij*} are centered within specialties so that they have a mean zero:

$$d_{ij} = d'_{ij*} - d'_{*j} \quad . \tag{1}$$

These d'_{ij} are on the logarithmic scale and are the differences from the average service within a specialty. All the specialty-specific scales are to be translated to a common scale and the relationships between the services within a specialty d_{ij} should remain the same. The linkage procedure accomplishes this outcome by shifting all services within a specialty by a fixed distance b_j. In other words, b_j is the position the specialty-specific origin is shifted to on the common scale. The location of any service on the common scale is

[2]Dempster, Laird, and Rubin (1977).

$$d_{ij} + b_j .$$

On the nonlogarithmic scale, 10^{b_j} is the multiplicative factor that changes the specialty-specific work units into common-scale work units. If the specialty has smaller units than average, then $b_j < 0$; if the specialty has larger units than average, then $b_j > 0$. An analogy is if the common scale is feet and radiologists measure in inches and surgeons in yards, the radiology constant $10^{b_j} = 1/12$ and the surgical constant $10^{b_j} = 3$.

Linked services should be as close as possible on the common scale. Define the optimal location of a linked service i on the common scale to be a_j. Using the specialty-specific shift described above, the deviation of any linked service on the common scale from the optimal location on the common scale is

$$d_{ij} + b_j - a_i .$$

The optimal parameters a_i and b_j are those that minimize the set of deviations over all linked services. The a_i and b_j are estimated via weighted least squares. The reason for taking logarithms of the ratings is that the error distribution in the linear model is closer to normal and thus the usual distribution theory and associated inference tests may be used when the regression model is examined.

The observations are weighted inversely to their estimated variances s^2_{ij}. That is, deviations with small variances will have more effect on the fitting as their values are better known.

Generally, one does not use weights in least-squares fitting because a unique estimate of variance at each observation is not known. Usually, the variance is assumed to be the same for all observations. However, because the deviations are actually the averages taken over the approximately 100 physicians that were sampled, the variance for each specific service may be estimated by the standard error of the mean.[3]

In addition to weighting by the variances, the Harvard study used the iterative Tukey biweight procedure.[4] The first step in this procedure is a weighted least squares using the inverses of variance esti-

[3]The estimated variance for standard services s^2_{kj} was calculated separately, as described in Section 2.

[4]Mosteller and Tukey (1977).

mates as weights. The residuals are used to reweight the observations and the new weighted least-squares problem is solved. This procedure is iterated until the biweights converge.[5] The idea behind this procedure is to give observations with large residuals less weight, as they are assumed to be outliers which have been poorly surveyed.

Although this procedure is statistically correct, it has an unanticipated effect if the biweight becomes zero: The link defined by the panel is in essence overruled and does not enter into the estimation of b_j. These "statistically eliminated" links should be scrutinized before accepting the results of the analysis, but such a scrutiny appears not to have been done.

The optimization problem is:

$$\min_{a_i,b_j} \Sigma[w_{ij}(d_{ij} + b_j - a_i)^2]/s^2_{ij} \qquad (2)$$

where the summation is taken over all specialties j and, for each specialty, over all linked services i. The minimization is done under the constraints that linked services have the same a_i and the mean of the b_j values is an arbitrary constant, taken in Phase II to be 2.025 to compare Phase I and Phase II results. The Tukey biweights are w_{ij}.

The 275 links resulted in 550 observations in the regression, as every link constitutes two observations, one for each link direction. The number of link location parameters a_i is 275, one for every link. The number of specialty shift parameters b_j is 39, one for each specialty surveyed in each phase of the Harvard study.[6]

After the fitting has been done, the logarithm of the work of specific service is estimated as $d_{ij} + b_j$, where the d_{ij} is observed and the b_j is estimated. The a_i does not appear in this equation. The linked services have just pulled each specialty's scale into alignment and the effect is seen in the locations of all specialty j's surveyed services. Given the above definitions, the position of the particular service on the common scale is related to the position of the particular service on the specialty-specific scale by a shift of b_j.

---

[5]In the Phase II Harvard study and our own calculations below, three Tukey biweight steps were taken.

[6]Although the number of distinct surveyed specialties was 32, seven specialties were surveyed during both phases. Two specialty shift parameters, one for each survey, were calculated for these specialties. For ease of discussion, we will throughout the rest of this section call the 39 survey sessions "specialties."

DUPLICATING THE HARVARD LINKAGE PROCEDURE

To fully understand the Harvard linkage procedure, we attempted to replicate it. Because only the means and standard deviations of services over physicians were available to us, we could not use, much less validate, the estimation-maximization averaging of the physician-level data. That is, our raw data were the d'_{ij*} values.[7]

In the previous subsection, the d'_{ij*} were discussed as if all were the same type of work. In actual fact, Phase II links could be among three different types of work: intra-service work, total work, and work per unit time (intensity). This resulted in four types of links used in Phase II:

- Links from intra-service work to intra-service work,
- Links from total work to total work,
- Links from total work to intra-service work, and
- Links from intensity to intensity.

For 1,126 surveyed services, data were available for intra-service work, total work, and intra-service time physician-level means, and associated standard errors and numbers of physicians surveyed for each type of work. One specialty, ophthalmology in Phase I, did not have estimated standard errors; we instead used the average standard error for all ophthalmological services in Phase II as a surrogate. After some exploration, we discovered that the Harvard study variance estimates were multiplied by the number of physicians surveyed for each service. We did likewise, although this multiplication gives less weight in the regression to those services that were more widely surveyed. We will refer to these weighted variances as the Harvard variances throughout the rest of this section.

We wanted to compare our estimated specialty shift parameters b_j with those reported by Harvard. Thus, we needed to first center the work logarithms d'_{ij*} to have mean zero as in Equation (1). In general, we subtracted the intra-service work mean within specialty from both the intra-service work values and the total work values. For three specialties,[8] we used the total work specialty mean, as intra-

[7]Recall that these d'_{ij} values are logarithms. Throughout this subsection, all calculations will be based on the logarithm of work, not on work itself. To help the text flow more smoothly, we will omit this specification most of the time.

[8]Nuclear medicine, radiation oncology, and pathology/Phase II.

service work values were not reported. For intensity links, we divided the centered intra-service work value by the intra-service time reported.

Given these centered work values, we first fit the model using weighted least squares, with the weights equal to the inverse of the Harvard variances. Then we took three Tukey steps, iterating new weight values at each step. Table 5 compares the Harvard reported value from the Phase II final report and our result, using the same degree of accuracy as Harvard reported. The final column of the table expresses the differences between the two linkage calculations in terms of the percentage change in specialty work values. That is, if Δ_j is the difference on the logarithmic scale between our result and the Harvard result, then $100[10^{\Delta j} - 1]$ expresses this difference in terms of a percentage increase or decrease in the work values for a specialty j. This value, which we call "percent change," is given in the last column of Table 5. This percentage difference is a measure of the change that would result in physician work (and hence payment) in adopting an alternative linkage procedure. For example, in Table 5, the Δ for plastic surgery is 0.2405. This translates to a percentage change of 5.69, which means that if our calculations were adopted instead of those of the Harvard group, the work value of all services measured and extrapolated from the plastic surgery survey would be increased by 5.69 percent.

In general, our results are within 10 percent of the Harvard results except for the three specialties, dermatology/II, ophthalmology/I, and orthopedic surgery/II. These differences could be due to the problems we had duplicating the Harvard centering procedure and the lack of standard errors for ophthalmology/I.

A NEW LOOK AT LINKAGE

Our examination of the Harvard linkage procedure revealed certain troublesome choices and simplifying assumptions. We therefore consider an alternative to their methodology, partly to determine how sensitive the results were to the linkage approach. In particular, our proposed alternative takes into account the following philosophical tenets that we consider important:

- If the linked services are stated to have equivalent amounts of work, then the RVW scale should reflect this equivalence, and the

Table 5

Comparison of Harvard Phase II Linkage and Our Replication Attempt

Specialty	Orig. b_j	RAND b_j	Δ_j	Percent Change
Allergy/immunology	1.8404	1.8590	0.0186	4.36
Anesthesiology	2.1921	2.2001	0.0080	1.85
Cardiology	2.1280	2.1213	–0.0068	–1.54
Dermatology/I	1.6385	1.6463	0.0078	1.80
Dermatology/II	1.9775	1.8737	–0.1038	–21.27
Emergency medicine	1.9367	1.9405	0.0038	0.87
Family practice	1.7794	1.7873	0.0078	1.82
Gastroenterology	2.1589	2.1860	0.0270	6.43
General surgery/I	2.2061	2.2309	0.0248	5.86
General surgery/II	2.4432	2.4603	0.0170	4.00
Hematology/oncology	1.9284	1.9163	–0.0122	–2.76
Infectious diseases	1.9255	1.9436	0.0180	4.24
Internal medicine/I	1.7579	1.7633	0.0054	1.24
Internal medicine/II	1.7988	1.8054	0.0066	1.52
Maxillofacial surgery	2.2449	2.2537	0.0088	2.04
Nephrology	2.0532	2.0456	–0.0076	–1.75
Neurology	1.8253	1.8005	–0.0248	–5.56
Neurosurgery	2.7556	2.7373	–0.0184	–4.14
Nuclear medicine	1.8806	1.8461	–0.0346	–7.65
Obstetrics/gynecology	2.0722	2.0855	0.0132	3.10
Ophthalmology/I	2.1181	2.1433	0.0252	5.96
Ophthalmology/II	2.1014	2.1748	0.0734	18.40
Orthopedic surgery/I	2.0714	2.0543	–0.0172	–3.87
Orthopedic surgery/II	2.4918	2.3484	–0.1434	–28.13
Osteopathy	1.8000	1.7641	–0.0360	–7.94
Otolaryngology	2.2746	2.2932	0.0186	4.36
Pathology/I	1.6191	1.6460	0.0268	6.38
Pathology/II	1.7571	1.7124	–0.0448	–9.79
Pediatrics	1.6741	1.6749	0.0008	0.17
Physical and rehab.	1.8874	1.9246	0.0372	8.93
Plastic surgery	2.3746	2.3987	0.0240	5.69
Pulmonary medicine	1.8895	1.8913	0.0018	0.40
Psychiatry/I	2.1026	2.1205	0.0178	4.20
Psychiatry/II	2.1094	2.0736	–0.0358	–7.92
Radiology	1.6811	1.7172	0.0360	8.66
Rheumatology	1.7105	1.7399	0.0294	6.99
Radiation oncology	2.0365	2.0227	–0.0138	–3.14
Thoracic surgery	2.4752	2.4986	0.0234	5.52
Urology	2.2467	2.2640	0.0173	4.05

NOTE: A roman numeral following a specialty refers to the phase of the Harvard study.

a_i values calculated in the linkage procedure should *be* the common scale values. In the current Harvard method, the a_i are ignored, so linked services, though they are claimed to be equivalent, can and do have different work values.[9]

- Again, if equivalence means equivalence, then links should be transitive. Thus, if service A entails the same work as service B and service B entails the same work as service C, then services A and C should also be equivalent and linked. The Harvard group adopted a fuzzier definition of equivalence and did not assume transitivity.
- The Medicare Fee Schedule mandates a single RVW for each CPT code. This is an implicit statement that the same CPT code represents, on average, the same work across specialties and constitutes an implicit link. These links should also be used to form the common scale.[10]
- Finally, if the a_i values resulting from the linkage calculations are taken to be work values on the common scale, the surveyed work values not linked should maintain as close a relationship as possible to the new a_i values as they did to the originally surveyed d_{ij} values.

These principles, taken together, result in what we call the *perturbation minimization* procedure to express work on a common scale. The primary difference between this new methodology and the Harvard approach is that the former incorporates the equivalence of services fully and directly into the optimization algorithm. Equivalence is incorporated fully by adding the link transitivity requirement and same CPT code links. Equivalence is incorporated directly by the fact that our procedure yields the final RVWs, thereby eliminating the additional averaging step needed by the Harvard group after their optimization. By eliminating the averaging, however, we cannot main-

---

[9]As we explained in Section 2, this necessitates the additional step of deriving a common work value for different vignettes carrying a common CPT code, both within a specialty and across specialties.

[10]A problem arises because multiple vignettes *within* a specialty can have the same CPT code. For example, a large number of "50 minute hour" office visits in psychiatry are coded 90844, even it is obvious that different types of visits involve different amounts of work. But the intent of the RBRVS is not to provide work value for all physician services but to provide a valid *average* work value for all physician services billed under a particular CPT code. Thus, within a specialty, an average of surveyed values for common CPT codes is an estimate of work, which can be linked to similar averages of other specialties.

tain the surveyed relationships between service work values within a specialty. In effect, we posit that the inter-specialty equivalencies defined by links are free of measurement error, and therefore the surveyed intra-specialty relationships must be adjusted relative to the links.[11] Our new procedure consists of two discrete steps, generating a different set of links and determining the optimal work values via a new linkage procedure. After describing this new methodology and its results, we discuss their implications.

Generate a Different Set of Links

The first step is to define the set of services to be used in the links. We began with our link set equal to the Harvard link set. We then changed this link set in three ways.

1. Drop Intensity Links. The intensity links seem contradictory given that the Harvard approach justified magnitude estimation because work was a varying and not necessarily linear function of time, intensity, technical ability, and mental judgment. Thus, we deleted these intensity links from our link set, leaving all Harvard intra-service, total, and mixed links. This resulted in the loss of 32 of the 275 original Harvard links.

2. Generate Common CPT Code Links. We expanded our link set so that all same CPT code surveyed services were linked; thus, they were linked across specialties. In several cases, a CPT code was surveyed more than once per specialty, albeit with different vignettes. We did not link these same CPT code services within a specialty. Instead, we first formed a new "service" whose work value was the weighted (by number of survey respondents) average of the work values of services within that specialty with the same CPT code. The standard errors were calculated in the usual manner for a weighted mean. This averaging produced 83 new services, bringing the total number of services to 1,209.

Any original Harvard links between two services with the same CPT code were left intact. However, we did not distinguish among linked and unlinked services in creating the common-CPT artificial services. Thus, two vignettes sharing a common CPT code but from different specialties might be used twice in the linkage procedure: once if they

[11] We do not defend this assumption as true in an absolute sense but offer it as part of a set of assumptions with as much claim to validity as the set of assumptions adopted by the Harvard group.

were paired as an original Harvard vignette link and once as part of their contributions to a multi-vignette common-CPT link.

3. Make Linkages Transitive. The Harvard links were not necessarily transitive, a property that we believed essential to ensuring the fairness of the procedure and its acceptability. Therefore, we expanded links to create transitive link subsets and defined the members of each associated *orbit* to consist of the interlinked services, using a term borrowed from the modern algebra literature.[12]

An orbit is an inclusive set of services that are linked together transitively. For example, if service A is linked to service B, and service B is linked to service C, then service A must be linked to service C to ensure transitivity. If services A, B, and C are linked thus and are not linked to any other services, then (A, B, C) forms an orbit with three associated links between its member services. If any of the services are linked to other services, new links are added to ensure transitivity and the orbit becomes larger. For example, if A is also linked to D, then links from B to D and from C to D are added and the orbit consists of (A, B, C, D) with six associated links.

Each orbit has an associated a_i, which we call an orbit location parameter. Implicitly, we changed the Harvard optimization constraint in Equation (2) so that all members of an orbit o must have the orbit location parameter a_o.

4. Drop EM Service Links. After averaging, forming the new same CPT code links and making all links transitive, we had 172 orbits made up of 13,102 links. The main reason for this large number of links was that several CPT codes appear in almost every specialty. Given our same CPT code link approach, these specialties all become linked, producing several large orbits. If a Harvard vignette link happens to fall into the same orbit, the orbit must expand to include that service and all services linked to it to satisfy the transitivity requirement. An orbit can encompass many services through this expansion process. In particular, the three largest orbits had approximately 8,000, 3,000, and 1,200 links.

The CPT codes that produce these large orbits are EM codes (CPTs 90000 through 90699), which tend to appear in almost all specialties. Since the influence of a particular code on the optimization results increases monotonically with the number of links the service appears in, these EM codes have a large influence and tend to swamp the effect of other codes in the linkage process. General dissatisfaction

[12]Gilbert (1976).

with the current use of CPT codes for EM services has been expressed,[13] and their surveyed work values are considered to be uncertain. Given this uncertainty, we decided to drop EM codes from consideration for common-CPT linkage so that they would not be allowed to have a major effect on RBRVS calculations. If an EM code was linked in the Harvard set, we retained this vignette link in our set. The resulting new number of orbits was 208, with 638 links.

Table 6 shows the vignette and common-CPT code links by specialty. Each link appears twice in this table, since it is composed of two services in different specialties. Thus, the total number of entries in this table is 1,276. The number of specialties increases from 39 to 42 because of the addition of three subspecialties, ophthalmology/corneal procedures, ophthalmology/glaucoma procedures, and child psychiatry, which were separately surveyed but did not appear in the vignette link set. Because these services have some common-CPT links, they now appear in the linkage procedure. Table 6 shows that some specialties, for example ophthalmology/II, have large increases in the number of links.

Before describing the new linkage procedure, evaluation of the effect of the new expanded, transitive link set alone on the results is warranted. Table 7 shows the specialty shift parameters that result from the Harvard linkage procedure using the new link set. We compare these results to our own replication of the Harvard results rather than the original Harvard results. This comparison is made because our two sets of results are based on the same data assumptions and thus provide a fairer comparison of the consequences of the new link set.

The largest change is in anesthesiology, which loses almost 75 percent. Dermatology/I, ophthalmology/I, orthopedic surgery/II, and pediatrics each gain over 25 percent whereas emergency medicine and physical and rehabilitative services suffer decreases.

A New Linkage Procedure

In the Harvard linkage procedure, after the specialty shift parameters b_j and the orbit location parameters a_i are estimated by the

[13]Lasker, Marquis, and Morrow (1991); Physician Payment Review Commission (1991). This dissatisfaction led to the replacement of the EM codes by a new set (numbered 99200–99499) in the 1992 version of CPT.

Table 6

Number of Vignette and Common-CPT Code Links

Specialty	Vignette	Common CPT	Total
Allergy/immunology	6	7	13
Anesthesiology	5	10	15
Cardiology	7	14	21
Dermatology/I	5	12	17
Dermatology/II	9	26	35
Emergency medicine	6	21	27
Family practice	29	21	50
Gastroenterology	12	15	27
General surgery/I	32	36	68
General surgery/II	21	16	37
Hematology/oncology	12	11	23
Infectious diseases	5	19	24
Internal medicine/I	32	37	69
Internal medicine/II	39	39	69
Maxillofacial surgery	4	4	8
Nephrology	11	8	19
Neurology	12	13	25
Neurosurgery	11	15	27
Nuclear medicine	4	2	6
Obstetrics/gynecology	11	13	24
Ophthalmology/I	11	21	32
Ophthalmology/II	17	70	87
Ophthalmology/cornea		44	44
Ophthalmology/glaucoma		39	39
Orthopedic surgery/I	17	19	36
Orthopedic surgery/II	19	13	32
Osteopathy	11	9	20
Otolaryngology	12	12	24
Pathology/I	4	20	24
Pathology/II	4	26	30
Child psychiatry		13	13
Pediatrics	8	18	26
Physical and rehab.	9	22	31
Plastic surgery	17	13	30
Pulmonary medicine	15	33	48
Psychiatry/I	10	15	25
Psychiatry/II	7	15	22
Radiology	8	6	14
Rheumatology	15	16	31
Radiation oncology	6	5	11
Thoracic surgery	10	20	30
Urology	13	11	24

NOTE: A roman numeral following a specialty refers to the phase of the Harvard study. In Phase II, separate surveys were conducted for general ophthalmology, procedures involving the cornea, and procedures related to glaucoma.

Table 7

Comparison of the New Link Set Results to the Harvard Link Set Results

Specialty	New Link Set b_j	RAND b_j	Δ_i	Percent Change
Allergy/immunology	1.9468	1.8590	0.0878	22.40
Anesthesiology	1.6090	2.2001	–0.5911	–74.36
Cardiology	2.1035	2.1212	–0.0178	–4.01
Dermatology/I	1.7718	1.6463	0.1255	33.51
Dermatology/II	1.8882	1.8737	0.0145	3.41
Emergency medicine	1.8537	1.9405	–0.0868	–18.11
Family practice	1.7603	1.7872	–0.0270	–6.02
Gastroenterology	2.1766	2.1860	–0.0094	–2.14
General surgery/I	2.2253	2.2309	–0.0055	–1.26
General surgery/II	2.4464	2.4603	–0.0139	–3.15
Hematology/oncology	1.9180	1.9163	0.0018	0.41
Infectious diseases	1.8883	1.9435	–0.0552	–11.94
Internal medicine/I	1.7438	1.7633	–0.0195	–4.39
Internal medicine/II	1.7833	1.8054	–0.0221	–4.96
Maxillofacial surgery	2.2715	2.2537	0.0178	4.19
Nephrology	2.0381	2.0456	–0.0075	–1.71
Neurology	1.7910	1.8005	–0.0095	–2.16
Neurosurgery	2.7877	2.7372	0.0504	12.31
Nuclear medicine	1.8610	1.8461	0.0150	3.50
Obstetrics/gynecology	2.0358	2.0855	–0.0497	–10.82
Ophthalmology/I	2.2457	2.1432	0.1024	26.60
Ophthalmology/II	2.1416	2.1748	–0.0332	–7.36
Ophthalmology/cornea	2.1768			
Ophthalmology/glaucoma	2.0546			
Orthopedic surgery/I	2.1231	2.0542	0.0689	17.19
Orthopedic surgery/II	2.4611	2.3483	0.1128	29.65
Osteopathy	1.7701	1.7640	0.0061	1.40
Otolaryngology	2.2493	2.2931	–0.0438	–9.60
Pathology/I	1.7207	1.6459	0.0748	18.80
Pathology/II	1.7009	1.7124	–0.0114	–2.59
Child psychiatry	2.0906			
Pediatrics	1.7734	1.6748	0.0986	25.48
Physical and rehab.	1.8512	1.9246	–0.0733	–15.53
Plastic surgery	2.3496	2.3987	–0.0491	–10.68
Pulmonary medicine	1.8773	1.8913	–0.0140	–3.17
Psychiatry/I	2.1653	2.1205	0.0449	10.88
Psychiatry/II	2.0471	2.0736	–0.0265	–5.91
Radiology	1.7943	1.7171	0.0772	19.44
Rheumatology	1.7831	1.7398	0.0433	10.49
Radiation oncology	2.0393	2.0226	0.0167	3.91
Thoracic surgery	2.5000	2.4985	0.0015	0.34
Urology	2.2223	2.2639	–0.0416	–9.14

NOTE: A roman numeral following a specialty refers to the phase of the Harvard study. In Phase II, separate surveys were conducted for general ophthalmology, procedures involving the cornea, and procedures related to glaucoma.

least-squares procedure, the location of an unlinked service whose centered specialty-specific work value was d_{ij} is just $d_{ij} + b_j$. This approach maintains the distance between services within specialty, that is, the intra-specialty service relationships and the entire specialty-specific scale is just shifted as a body. However, services linked across specialties may not have the same work value after linkage because the a_i are not used to assign linked service work values on the common scale.

Our proposed alternative requires that all linked services within an orbit have the same work value on the common scale. That is, members of an orbit all have work value a_i. At the same time, we would like the optimization to ensure that the distances between services within a specialty stay as close as possible to the original surveyed distances. In essence, after redefining the linked service values to be a_i, we seek to adjust the unlinked d_{ij} values to preserve as much as possible their relationships to other services within their specialty. The name of our new procedure, perturbation minimization, describes this goal.

For services h and i within specialty j, let the surveyed distance be[14]

$$\delta_{hi,j} = d_{ij} - d_{hj} .$$

Let the location of an orbit o on the common scale be a_o and the location of services h and i be a_h and a_i. Then the new optimization problem is

$$\min_{a_o,a_h,a_i} \Sigma\Sigma\Sigma (a_i - a_h - \delta_{hi,j})^2/s^2_{hi,j} . \qquad (3)$$

The first summation is over all specialties j and the second and third are over all pairs of services (h,i) within a particular specialty with no double counting allowed. The objective function is minimized under the constraint that all member services of orbit o have a fixed value a_o. Each term is weighted by the inverse of its estimated variance $s^2_{hi,j}$, which is calculated from the surveyed variances of the d_{ij}. We do not use the Tukey biweight method.

The key difference between the optimizations in Equation (2) and Equation (3) is that all services are involved in the new optimization Equation (3) and services can move within specialties. We look at dif-

[14]If service k is the standard service for the specialty, i.e., h = k, then $\delta_{ki,j} = d_{ij}$. This convenience facilitates calculating other differences through the shortcut of $\delta_{hi,j} = d_{ij} - d_{hj}$.

ferences not only between the standard service and every other service in a specialty but also between all pairs of services within a specialty. We force linked services to have equal values for all members of an orbit and then try to minimize the effect on the relationships between services within specialties. We seek to minimize the perturbations within specialty scales that result from the required equality of all services within an orbit.

Though Equation (3) remains a least-squares problem, the number of parameters is large and the optimization problem could prove difficult to solve because of its size. A reasonable alternative[15] is a two-stage optimization. The first stage is to do the Harvard optimization of Equation (2) with the new link set and orbit constraints as was done for Table 7. After this stage, the work values of all members of the same orbit are set equal to their associated a_o. Then for each specialty j, we attempt to minimize the effect on the relationships between services within the specialty by solving the inner optimization from Equation (3):

$$\min_{a_h,a_i} \sum\sum (a_i - a_h - \delta_{hi,j})^2/s^2_{hi,j}$$

with the constraint that all linked services have $a_i = a_o$ for the appropriate orbit location parameter estimated in stage one. This alternative is equivalent philosophically to the full optimization in that it forces linked services to have the same assigned work value and it seeks to assign work values that minimize the distortion of surveyed distances between services within specialties that results from linkage.

Results of the Perturbation Minimization Procedure

The new perturbation minimization procedure does not require any calculations after it has been completed. In contrast, the Harvard procedure requires an averaging step afterward to ensure that linked services have the same work value and a CPT has a single work value. Thus, the results of our new procedure must be compared to Harvard's final reported results. Because of the uncertainty of the EM CPT codes and our subsequent decision not to use them to generate new links, we do not include them in our comparisons.

[15]We thank Grace Carter for her considerable assistance in developing this alternative.

The Harvard values used for comparison are the Phase III results reported in February 1991 to HCFA for surveyed non-EM services. These work values and our new values were standardized to have the same overall mean on the logarithmic scale before comparison, just as the different phase results were standardized by the Harvard group. Since the perturbation minimization procedure may shift work values within a specialty by varying amounts, comparison must be made between our and Harvard's results for each service individually. In contrast, earlier comparisons in Tables 5 and 7 were made by specialty, as the specialty scales shift rigidly with all services within a specialty moving a fixed amount.

The differences on the logarithmic scale were assessed graphically via a histogram for each specialty. In general, these histograms were unimodal and symmetric. The comparisons are summarized in Table 8 and in Figure 1.

Table 8 shows the average percentage difference between the work value calculated by the alternative linkage procedure and that calculated by the Harvard linkage procedure by specialty. Seven hundred services were compared across the 42 specialties to calculate these means. These 700 services corresponded to 522 CPT codes, as some codes appeared in more than one specialty. Of these 522, Harvard did not publish work values for 46. In Table 8, the mean of the differences between our work values and the Harvard work values on the logarithmic scale is given for each specialty in the column labeled Δ_j. The middle column shows the standard deviation of the differences on the logarithmic scale. The final column shows the percentage change in the work values on the nonlogarithmic scale that corresponds to the mean difference.

Figure 1 shows a histogram of the individual percentage differences for the 476 surveyed CPT codes for which Harvard published work values. Although most of the RAND RVWs are within 15 percent of the original values, one-fifth of the CPT codes have larger discrepancies.

In general, though, the percentage changes are smaller than those shown in Table 7, which indicates that the second stage of the perturbation minimization optimization (Equation (3)) has a similar effect to the Harvard averaging procedure. This similarity makes sense, since the new procedure sought to take into account the averaging goals automatically by including same CPT code links and by forcing members of an orbit to have the same work value. As before, the specialty with the largest change is anesthesiology with a loss of over 70

Table 8

Comparison of Harvard Phase III Final Work Values and Our Perturbation Minimization Results

Specialty	Δ_j	S.D.	Percent Change
Allergy/immunology	0.054	0.078	13.16
Anesthesiology	–0.536	0.206	–70.88
Cardiology	0.002	0.066	0.37
Dermatology/I	0.050	0.069	12.31
Dermatology/II	0.037	0.108	8.97
Emergency medicine	–0.017	0.103	–3.76
Family practice	0.004	0.083	1.01
Gastroenterology	0.008	0.081	1.88
General surgery/I	0.006	0.258	1.50
General surgery/II	0.024	0.022	5.79
Hematology/oncology	0.014	0.040	3.41
Infectious diseases	0.026	0.046	6.41
Internal medicine/I	–0.009	0.086	–2.09
Internal medicine/II	0.048	0.074	11.62
Maxillofacial surgery	0.024	0.179	5.76
Nephrology	0.010	0.011	2.41
Neurology	–0.001	0.011	–0.13
Neurosurgery	0.070	0.063	17.41
Nuclear medicine	–0.013	0.014	–2.89
Obstetrics/gynecology	–0.007	0.109	–1.64
Ophthalmology/I	0.045	0.183	10.96
Ophthalmology/II	–0.018	0.160	–4.12
Ophthalmology/cornea	–0.023	0.175	–5.07
Ophthalmology/glaucoma	–0.027	0.103	–6.09
Orthopedic surgery/I	0.063	0.075	15.51
Orthopedic surgery/II	0.049	0.090	11.83
Osteopathy	–0.019	0.073	–4.38
Otolaryngology	0.024	0.060	5.62
Pathology/I	–0.039	0.046	–8.52
Pathology/II	–0.020	0.116	–4.50
Child psychiatry	0.114	0.256	30.12
Pediatrics	–0.019	0.092	–4.28
Physical and rehab.	–0.004	0.042	–0.96
Plastic surgery	0.030	0.057	7.15
Pulmonary medicine	0.041	0.036	9.99
Psychiatry/I	0.081	0.274	20.58
Psychiatry/II	–0.010	0.196	–2.22
Radiology	–0.021	0.029	–4.66
Rheumatology	0.048	0.059	11.76
Radiation oncology	0.073	0.214	18.30
Thoracic surgery	0.090	0.037	23.08
Urology	0.063	0.056	15.61

NOTE: A roman numeral following a specialty refers to the phase of the Harvard study. In Phase II, separate surveys were conducted for general ophthalmology, procedures involving the cornea, and procedures related to glaucoma.

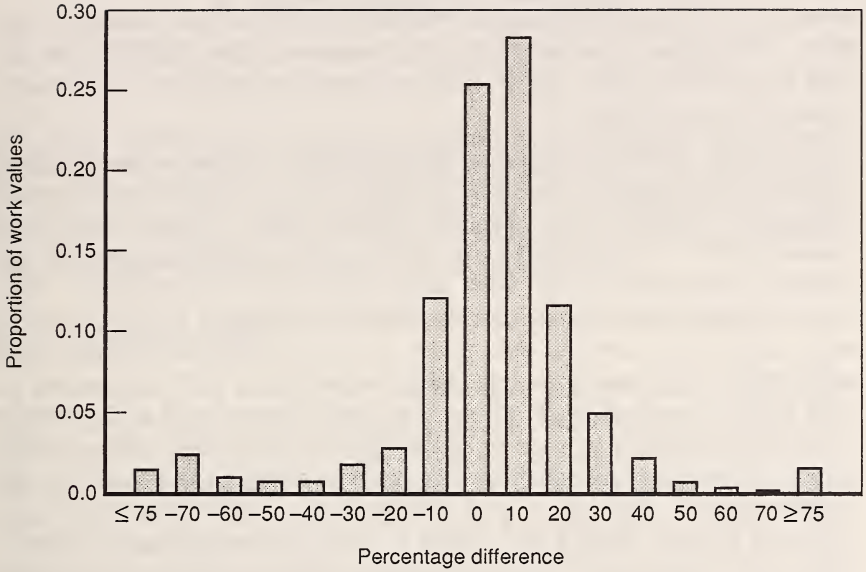

Figure 1—Histogram of Percentage Differences in Work Values, RAND vs. Harvard Linkage Procedure

percent. Neurosurgery, child psychiatry, psychiatry/I, radiation oncology, and thoracic surgery all have gains of over 15 percent. When considering these averages, the standard deviation of the differences must also be taken into account. A large standard deviation indicates that the percentage changes for services within a specialty varied widely.

Implications of the Alternative Methodology

The conclusion from our alternative linkage exercise is that the work values change considerably depending on the method used as evidenced in Table 8. This means that the results of a linkage procedure are sensitive to the assumptions underlying that procedure. Although our personal preference is for our set of assumptions over those of the Harvard group, our claim is only equal preference. We do not claim that our results are better than Harvard's, only that they are different.

Our conclusion is just a beginning. Without further sensitivity analysis to investigate the behavior of the results, for example with the

deletion or addition of links, the validity of the Harvard results is unclear. Given the considerable public response to the RBRVS, the future use of this or any other linkage procedure without much more research is not advised.

One possible criticism of our method is that a single parameter that embodies a particular specialty's shift to the common scale is not available. The Harvard method produced single values, the b_j discussed above. Our method may shift services within a specialty by different amounts, so no such sole parameter is available. The Harvard group used these specialty shift parameters to calculate the work values for services that were surveyed after the linkage was performed. The assumption in using a specialty's shift parameter to shift future surveyed work values onto the common scale is that these surveyed values are measured in the same units as the original ones. However, this assumption is violated in this study by the fact that the Harvard shift parameters are different for specialties that were surveyed in both Phase I and Phase II, such as dermatology (Table 5). Thus, this criticism of our alternative linkage procedure is not supported.

A key question in hindsight is: Why link at all? The need for linkage arose because physicians ranked their specialty's services on different scales which then had to be calibrated onto a common scale. In retrospect, that may not have been the most desirable strategy; establishing instead a common inter-service scale for all specialty surveys might have been preferable.

In any event, the issue is moot with respect to the present set of values. For the future, however, the lesson learned from our investigation is that any linkage procedure is sensitive to its assumptions and better avoided if possible. Because the published RVWs provide a common scale for any future measurements, we recommend that future surveys of physician work be based on that common scale, thereby eliminating any need for linkage. Even if some of the RVWs in the published set are questionable, the set is large enough and dense enough so that some of the values are consensually regarded as accurate and can form a reference value foundation for future measurement sessions.

5. RECOMMENDATIONS

The work reported here has concentrated on two aspects of obtaining relative work values for the Medicare Fee Schedule. We have looked at who should be surveyed for data to construct the RBRVS, how data should be collected, and how data collected from diverse specialities can be combined into a common scale of magnitude.

COLLECTING DATA

On the basis of our scrutiny of the Harvard study Phase II examination of alternative data-collection methods, in conjunction with our review of the psychological literature on individual- and collective-based decisionmaking, we recommend that any future magnitude estimation of work values be done using a group-based method that provides intermediate feedback to group members to permit them to adjust their individual numerical estimations. The two leading candidates for such a method are (1) a multiple-round mail survey with feedback on the distribution of responses between rounds (i.e., a Delphi process) and (2) a discussion panel preceded by a preliminary mail round. In terms of the validity of responses, we have no evidence that suggests any major differences between methods. The discussion method costs about 10 percent more than the Delphi method, largely because of the travel costs required to assemble the group.

With costs approximately equal and with no expectation that the two group methods will differ in results, the choice between them can be based on priorities not related to validity or expense. Because of the political sensitivity of the MFS, the Delphi method may be preferred because it surveys a larger respondent sample and thereby permits greater representativeness among geographical, experience, gender, racial, and other potential stratifying factors. Therefore, for *long-term* updating of the RBRVS, the Delphi method may be the preferred one. However, in response to *short-term* demands, where answers need to arrive in timely fashion, a discussion panel is much more easily assembled and is therefore probably the better option.

To answer the question of who should provide the data, the need for physicians experienced in the procedures is logically unassailable. However, the presence of experienced physicians need not imply the absence of others; the leavening effect of primary care physicians or other specialty representation lessens the opportunities for gaming

and helps keep measurements from different specialties aligned on a common scale. Because of the potential for conflict of interest, specialty societies should not have exclusive say in selecting respondents to Delphi surveys or participants in discussion panels. We recommend that the new HCFA universal provider file, from which physicians with the necessary experience in the target services can be identified, be used for these purposes. A possible role for the specialty society representatives might be as consultative experts to panels, discussing issues but not doing the ratings themselves. In this way, their expertise informs the panels but does not preordain the results.

CREATING A COMMON SCALE

Our examination of the Harvard study's linkage procedures has unearthed a number of possible technical problems and conceptual ambiguities. The technical problems, which include choice of individual or group standard error as a weighting factor in linkage, statistically eliminated links when biweights become zero, and some apparent inconsistencies in link choices, are all easily fixed. The majority of these problems do not involve changes of more than 5 percent in work values.

The conceptual ambiguities, however, call into question the validity of the published RBRVS. Our preliminary analysis shows that an alternative specification of the linkage procedure, which we believe to be at least as consistent with the philosophy and intent of the Harvard group's specification, produces adjustments that are more than trivially different. Linkage is demonstrably sensitive to underlying assumptions and until it is clearly understood through further investigation, it should not be used in future revisions of the MFS.

However, given that a set of RVWs based on a common scale of measurement has been published and that many of those work values appear to be acceptable to the medical profession, the need for links to move individual specialty work value estimates to a common scale may be obviated. If a set of reference values that contains frequently performed services across specialties and across the continuum of work values can be validated, then that set can be employed as a defining "ruler" for any future estimates of relative work.

AN OVERALL VIEW

Nobody claims that the RBRVS is perfect. It is not perfect now and it will never be perfect. However, hardly anybody claims that the system it replaces wasn't broken. The system of customary, reasonable,

and prevailing charges was too arbitrary and too uncertain to be relied upon as the basis for a rational payment policy. The RBRVS, even with all of its imperfections, represents an advance. Future work—even work that causes major changes in the RBRVS—can similarly be viewed as progress toward the objective of fair and consistent policies for Medicare physician payment.

Appendix
SUMMARY OF RBRVS DEVELOPMENT

SUMMARY OF RBRVS DEVELOPMENT

Harvard Phase I	Research Methods
1. Develop Relative Work Values (RWVs) for vignettes (i.e., services) measuring physician work provided to "average" patients	1. Conduct specialty-specific *telephone* surveys for 18 specialties (allergy and immunology; anesthesiology; dermatology; family practice; general surgery; internal medicine; obstetrics and gynecology; maxillofacial and oral surgery; ophthalmology; orthopedic surgery; otolaryngology; pathology; pediatrics; psychiatry; radiology; rheumatology; thoracic and cardiovascular surgery; and urology):
a. Identify ~23 services, including a "standard" service, within each specialty	a. Use 100 MDs from private practice and academic medicine organized into 14 Technical Consulting Groups (TCGs) to identify services for inclusion in telephone survey for each specialty. Services selected to cover 4 broad types of activity: (1) evaluation and management (EM); (2) invasive; (3) laboratory; and (4) imaging and pattern recognition.
b. Obtain work estimates relative to a standard service within each specialty	b. Select 185 MDs per specialty based on a stratified random sample from 1986 AMA Masterfile (except for maxillofacial and oral surgery), and conduct telephone survey with ~100 MDs per specialty. 3,164 eligible MDs contacted; 1,977 interviewed. Use *magnitude estimation* to obtain estimates of total *intra-service* work, plus 3 dimensions of work (mental effort and judgment; technical skill and physical effort; and stress) compared to a "standard" service, where work for the standard service is defined as 100 for these 4 variables. Also, obtain time estimates, defined as encounter time, for all services, including standard service. Use estimation-maximization algorithm to fill in nonresponse missing values and values excluded as outliers. This step yielded direct estimates of intra-service work and time for 389 of the 407 vignettes (after 2 were dropped), corresponding to 372 unique services and 239 unique CPT codes.
2. Develop cross-specialty links between 18 specialties to produce relative work values on a common scale	2. a. Use multi-specialty panel of 24 MDs to identify "same" and "equivalent" services to serve as links, based on *intra-service work*. 159 potential linkages identified originally, reduced to 132 after first meeting of panel, reduced to 103 after eliminating nonsurveyed specialties, reduced to 75 after eliminating links with greater than 25 percent difference in average time between services. Final

		number increased to 82 (40 "same," 42 "equivalent") after performing cluster analysis on time and work to identify additional potential links that were approved by the physician panel.
		b. Perform a weighted least-squares analysis using *intra-service work* for linked services to determine *optimal, compromise* scaling factors (b_j's) within each specialty that permit realignment of all specialty-specific work ratings onto a common scale.
		c. For vignettes within a specialty that map into the same CPT code, calculate final work value as the simple arithmetic mean of each vignette. For vignettes in different specialties that map into the same CPT code, calculate final work value on the common scale as the volume-weighted average (using Part B data) of work values on the realigned specialty-specific scales.
3.	Measure pre- and post-service work for selected services to develop estimates of total work	3. a. For 55 of 407 vignettes in original survey, obtain estimates of pre- and post-service *times*.
		b. Conduct a follow-up telephone survey among MDs who participated in original survey in 7 specialties (allergy and immunology; general surgery; internal medicine; maxillofacial and oral surgery; orthopedic surgery; pathology; and thoracic surgery) to obtain estimates of pre-, intra-, and post-service times. This step produced data for 121 services.
		c. Develop multivariate models, with pre- and post-service times as a function of intra-service work and intra-service time, to extrapolate to surveyed services with intra-service work values only. Use separate models for 8 classes of service (4 pre-service: office EM and imaging; hospital and emergency room EM; office invasive; and surgical center and hospital invasive; 4 post-service: office, hospital, and emergency room EM and imaging; office invasive; surgical center invasive; and hospital invasive).

		d.	For 144 EM services, calculate work intensity based on *intra-service* work per minute (W/T). For 17 services in the original survey, deflate pre- and post-service time estimates by 30 percent. Aggregate EM intensity values into 4 levels. Then, use cluster analysis to define 9 categories of services based on intra-service work. Then, multiply extrapolated pre- and post-service times for each category of service by appropriate level of EM intensity (level 1=2.16; 2=2.80; 3=3.44; 4=4.85) to obtain estimates of pre- and post-service work. Add these values of pre- and post-service work to intra-service work estimates to calculate *total work*.
4.	Extrapolate total work from surveyed services to other, nonsurveyed services within "families" of services	4. a.	Create "families" of CPT codes based on 4 broad categories of service (EM; invasive; imaging; and laboratory), grouped further by setting, body system, anatomic region, type of procedure, etc.
		b.	Assign 372 surveyed services to families. Subdivide families with more than one surveyed service so that each family has only one surveyed service. This surveyed service serves as the "benchmark."
		c.	For 293 families (excluding anesthesiology), calculate ratio of average *submitted* charges for each nonsurveyed service within a family to average *submitted* charges for the "benchmark" service, using pooled 1986 Medicare Part B data and HIAA data from 1986/1987. Then, multiply this ratio by the total work value for the "benchmark" service to obtain an extrapolated work value for each nonsurveyed service. For EM services, extrapolations were specialty-specific, because the same CPT codes had work ratings that varied widely across specialties. After extrapolation, Phase I produced *total work* values for ~1,400 services accounting for ~67 percent of total Part B allowed charges, and ~80 percent of allowed charges for surgical services.

Harvard Phase II			Research Methods
1.	a. Survey 15 additional specialties not included in Phase I	1. a.	For 15 additional specialties (cardiology; emergency medicine; gastroenterology; hematology/oncology; infectious disease; nephrology; neurology; neurosurgery; nuclear medicine; osteopathy; physical and rehabilitative medicine; plastic surgery; pulmonary medicine; and radiation oncology) obtain estimates of *intra-service* work and time, plus the 3 dimensions of work, using *telephone* survey methodology from Phase I. For selected services, obtain estimates of *total* work

		and time. 2,607 eligible MDs contacted; 1,877 interviewed. Sample drawn from 1988 AMA Masterfile. Adjust MD-specific ratings of work for different "perceptions" of the standard service (i.e., calculate an adjustment factor that yields a geometric mean rating of 100 across all services).
	b. Resurvey 3 specialties from Phase I due to high volume of services performed	b. Obtain estimates of *intra-service* work and time *only* for additional services in general surgery, internal medicine, and orthopedic surgery using *telephone* survey methodology from Phase I. Also, obtain estimates of *total* work and time for selected services in internal medicine.
	c. Resurvey 4 specialties from Phase I to refine original estimates	c. Obtain additional estimates of *intra-service* work and time for dermatology, ophthalmology, pathology, and psychiatry using *telephone* survey methodology from Phase I. Also, obtain estimates of *total* work and time for selected services in ophthalmology and psychiatry. For 7 specialties resurveyed, 1,070 eligible MDs contacted, 844 interviewed.
		d. Adjust MD ratings of work for different "perceptions" of the standard service.
		e. For all 22 specialties combined in Phase II, estimates obtained for 753 vignettes involving 409 unique CPT codes.
2. Refine cross-specialty linkages		2. a. Use multi-specialty panels consisting of subsets from 26 specialties to identify "same" and "equivalent" services to serve as linkages. Use 4 types of linkages, based on: (1) intra-service work; (2) total work; (3) intensity; or (4) intra/total. In 5 multi-specialty panel meetings, panelists identified 193 pairs of services from 362 potential links.
		b. Use two multi-specialty panels of *salaried* MDs (one from staff-model HMOs, one from VA hospitals) from 9 specialties to establish an independent set of links.
		c. Link 33 specialties (15 from Phase II, 3 from Phase I resurveyed in Phase II, 4 resurveyed as special studies in Phase II, and 11 from Phase I). For specialties surveyed in both Phase I and II, use *both* Phase I and Phase II results. For 5 specialties from Phase I (urology, otolaryngology, radiology, thoracic and cardiovascular surgery, and rheumatology), develop additional linkages. For remaining Phase I specialties, use original linkages. Final number of linkages is 275 (82 from Phase I, 193 from Phase II).

d. Refine Phase I weighted-least squares method to include: (1) imputed standard deviations for standard services used as links, and (2) Tukey "bi-squared" weights to reduce the impact of extreme values.

e. Allow nontransitive linkages, i.e., A=B, B=C, but A ne C.

f. For vignettes within a specialty that map into the same CPT code, calculate final work value as the simple arithmetic mean of each vignette. For vignettes in different specialties that map into the same CPT code, calculate final work value on the common scale as the volume-weighted average (using **Part** B data) of work values on the realigned specialty-specific scales.

3. Refine estimates of pre- and post-service work for EM and surgical services, and refine definition of intra-service work

3. Define intra-service work as: (1) face-to-face encounter time for office visits; (2) time spent on floor for a hospital visit; (3) skin-to-skin time for a surgical procedure; and (4) intra-, pre-, and post-service work combined for laboratory and imaging services.

For *EM services*:

a. As part of Phase II telephone survey, obtain estimates of *aggregate* time spent on pre-, intra-, and post-service work *during a typical week*. Also, obtain estimates of *total* work and time for 99 services categorized into 4 types: (1) office visits; (2) consultations; (3) initial hospital visit; and (4) other.

b. Validate findings for EM services using: (1) review by TCG panelists, (2) comparison of work intensity (i.e., W/T) during the pre- and post-service periods with work intensity during the intra-service, and (3) comparison of survey results with aggregate data on amount of time spent by MDs in different activities. Adjust survey estimates of pre- and post-service work so that W/T for pre- and post-service work is 82 percent of W/T for *intra-service* work for each of the 4 types of EM services.

For *invasive services*, obtain additional estimates of pre- and post-service *work and time* for 109 selected services, and use regression to determine pre- and post-service *times* for all other surveyed services (i.e., for all other services with estimates of only *intra-service work and time*.) Also, calculate work *intensity (W/T)* for each component of pre- and post-service work for 31 surveyed services, then multiply extrapolated pre- and post-service times by W/T to obtain pre- and post-service work.

- Define pre- and post-service work as 8 components: (1) initial consultation; (2) hospital admission work-up; (3) pre-operative evaluation; (4) other pre-op work; (5) post-op follow-up on day of surgery; (6) follow-up visits in ICU after day of surgery; (7) follow-up visits in acute care unit after day of surgery; (8) post-hospital follow-up visits within 90 days of surgery.

- Obtain additional estimates of pre- and post-service *time* from: (1) 6 specialties for 26 services as part of *telephone* survey; (2) expert panel in general surgery for 30 procedures using a separate *mail* survey; and (3) TCG panelists in all surgical specialties for 111 procedures using a separate *mail* survey. This produced estimates of pre-operative evaluation time, same-day post-service time, and hospital post-service time for 109 invasive services.

- From national *telephone* survey, obtain estimates of work and time for 31 EM services (6 hospital admissions, 11 follow-up hospital visits, and 14 follow-up office visits) that occur before or after invasive procedures. Use these estimates to calculate (W/T) for each service.

- Use median inpatient length of stay to validate estimates of number of post-operative hospital visits obtained from survey.

- Obtain estimates of number of post-discharge office visits for each surgical procedure from specialty societies solicited by PPRC. Compare these estimates with survey and with estimates from TCG panelists.

- Use 6 regression models to estimate three components of pre- and post-service *time* as a function of intra-service work, intra-service time, hospital median LOS, and category of surgical service. The 6 models represent 3 components of peri-service time: (1) pre-op (component 3); (2) same day post-op (components 5+6+7); and (3) office follow-up (component 8); separated according to inpatient and ambulatory surgical setting. For other pre-op work (component 4), assign fixed value of 0, 15, or 25 minutes, depending on procedure and setting.

- Assign (W/T) values for each component of the service: (1) 2.2 for pre-op; (2) 0.8 for other pre-op; (3) 3.0 for same-day post-op; and (4) 2.5 for office follow-up.
- For each component, multiply predicted time from regression model by intensity to obtain work, then sum across components to get *total peri-service work*. Because of uncertainty about global fee policy, group services into 3 categories: (1) invasive procedures, which include all components of work; (2) endoscopic procedures, which include only work performed on day of procedure; and (3) minor procedures, which exclude pre- and post-service work.

4.	Refine extrapolations to determine total work values for surgical services	4. a.	For 3 resurveyed specialties, increase number of services surveyed to reduce need for extrapolation for high-volume services.
		b.	Identify 3 catagories of extrapolation problems: (1) heterogeneous families; (2) low-volume codes; and (3) other.
		c.	Use TCGs to identify more homogeneous families based on similarity in: (1) technology, (2) setting, (3) specialty, and (4) "maturity" of service, i.e., how new is it. Exclude obsolete or vague CPT codes. Link families without "benchmarks" to related families to increase number of services for which extrapolations can be calculated.
		d.	Validate extrapolated values by comparing surveyed values with extrapolated values in families with more than one surveyed service. This involved 104 surgical services in 39 families.
		e.	Use updated 1988 BMAD data and HIAA data from March 1988 through February 1989. For each service, the extrapolation ratio is equal to the volume-weighted average of the average charge in each data base.
		f.	Use additional claims data from 3 states for low-volume services (i.e., with <200 claims in the BMAD/HIAA file).
		g.	Exclude extrapolated values lower than 0.25 or higher than 4.0 relative to the benchmark. Also, exclude dermatology and ophthalmology services.

		h. These steps produced *total work* values for 2,024 CPT codes (200 surveyed plus 1,824 extrapolated), accounting for ~84 percent of Part B allowed charges for surgical services.
5.	Evaluate structure of EM codes	5. Use regression models where total work is a function of: (1) intra-service time; (2) service site (office, hospital, nursing home, or ER); (3) service type (medical/preventive or consultative); and (4) type of patient (new or established). Estimate models with: (1) one intercept pooled across service sites and types; (2) different intercepts by type of service; (3) different intercepts by type and site of service; (4) different intercepts and coefficients by type and site of service (for 7 groups with sufficient observations).
6.	Develop "cross-walk" between current EM codes and potential new EM coding system	6. For medical specialties, identify person who actually codes office and hospital visits for MDs participating in survey. Then, obtain CPT code assignments from persons responsible for coding for 32 EM vignettes in 9 specialties. Compare codes assigned in survey with codes assigned by TCG panelists.
7.	Examine impact of patient age on estimates of intra-service and total work	7. Compare survey results for 3 pairs of services in 3 specialties (hematology/oncology, internal medicine, and physical medicine).
8.	Validate survey results	8. a. For general surgery, organize 3 MD panels and obtain "independent" estimates of intra-service work for the same 38 services included in the Phase II national survey, plus 17 nonsurveyed services, using different methods. Panel A had 11 surgeons who were surveyed using a 2-round Delphi process followed by 2 face-to-face meetings. Panel B had 19 surgeons who were surveyed using a 3-round Delphi process. Panel C consisted of 29 surgeons who participated in a 1-round mail survey of 25 services, including pre- and post-service work. Adjust MD-specific ratings of work for different "perceptions" of the standard service. Based on findings from these 3 panels, propose 1-round mail survey to update ratings of work after implementation of Medicare Fee Schedule, and to obtain additional work ratings in Phase III study.
		b. For 13 services in 3 specialties (general surgery, internal medicine, and orthopedic surgery), compare Phase I and Phase II survey results.

Harvard Phase III	Research Methods
	Use expert panels of 15 MDs per specialty. Conduct multiple one-round mail surveys containing 50 services per survey:
1. Gap-fill values for low-volume and new CPT code. (Phases I and II produced RWVs for 2,412 codes in 262 families, accounting for ~84 percent of Part B surgical claims. Phase III will address remaining 1,966 codes in 185 families, yielding RWVs for 4,378 surgical codes in 447 families.)	1. For 26 specialties (allergy and immunology cardiology; dermatology; emergency medicine; family practice; gastroenterology; general surgery; hematology; infectious disease; internal medicine; nephrology; neurology; neurosurgery; nuclear medicine; obstetrics and gynecology; oncology; ophthalmology; orthopedic surgery; otolaryngology; physical medicine; plastic surgery; pulmonary medicine; radiation therapy; radiology; thoracic and cardiovascular surgery; urology) (7 excluded are anesthesiology; oral surgery; osteopathy; pathology; pediatrics; psychiatry; rheumatology), panels will: a. Obtain estimates of *total* work for standard service and other high-volume services in each family, and *intra-service* work values for all remaining services within each family, as well as for new services and changing technologies or practice patterns. b. For allergy and immunology, obtain total work estimates for all CPT codes to avoid extrapolation problems.
2. Reexamine RWVs for all extrapolated codes	2. a. Attempt to obtain direct *total work* ratings for all extrapolated services from Phases I and II. b. Expert panels will not be able to develop direct estimates for all services by September 1991. Therefore, for 6 specialties (general surgery, orthopedic surgery, ophthalmology, urology, gastroenterology, and thoracic and cardiovascular surgery), use specialty-specific panels of 11 MDs to examine all extrapolated and "sensitive" values, and to identify values that are more than 30 percent different from the "reasonable" value. Then, ask specialty-specific expert panels to provide new ratings of work. c. Compare Phase I extrapolated values with Phase II surveyed values for services included in both phases.

3.	Estimate pre- and post-operative work values for surgical services to determine the total work under a global surgical service	3. a. Global services defined as 5 components: (1) pre-surgical EM; (2) other pre-surgical work; (3) post-operative follow-up on day of surgery; (4) follow-up visits in hospital after day of surgery; and (5) follow-up visits in office. b. Obtain direct estimates of work and time during the pre- and post-operative periods for ~300 additional surgical procedures, including the number, duration, and work values of visits before and after surgery. Compare estimates based on 5 components with direct estimates of total work and time for entire global service. Finally, obtain estimates of global service in two portions (e.g., before and after hospital discharge). c. Validate direct estimates from expert panels using estimates from national surveys and data on hospital LOS and on follow-up office visits. d. Refine multivariate models used to obtain pre- and post-service work estimates for surveyed procedures where only *intra-service* work was obtained. Possible refinements include hospital LOS, direct estimates of relative stress from national surveys, use of ICU, mortality rates, and complication rates. e. Use HMO data on pre- and post-service work. Each expert panel will follow the same "modified Delphi" small-group process, involving: • Instructions for the physician leader and project staff concerning the definitions and panel process, • Use of vignettes for CPT codes within families that represent a "typical" patient, • Initial round of mailed surveys, • Follow-up by mail and phone to assure return of all Round 1 surveys, • Compilation of Round 1 results by project staff, • Round 2 surveys, including results from Round 1, to obtain new estimate for *low-volume* services, and • Use of national surveys as "gold standards" to validate results from each small group.
4.	Improve codes for EM and invasive services under CPT	4. Using direct estimates of time and work for 185 EM codes obtained from 32 specialties in Phases I and II:

		a.	Compare work and time for the same CPT code across specialties.
		b.	Compare CPT codes used to code the same service across specialties.
		c.	Analyze range of work and time by type of EM service.
		d.	Analyze "levels" of service by specialty and type of service grouped by work and/or time.
5.	Link final RWVs with scales developed by the American College of Radiologists and the American Society of Anesthesiologists	5. a.	Use expert panel in radiology to obtain direct estimates of work for the *professional* component of radiological services.
		b.	Use OR time logs to develop average anesthesia times per procedure.

	PPRC	Research Methods
1.	Conduct Survey of Visits and Consultations	1. 339 MDs in 3 specialties (internal medicine, rheumatology, and urology) participated in study measuring time for 7 categories of EM services provided in the hospital and office during the day of the visit. The 7 categories were: (1) record review; (2) history and physical exam; (3) counseling; (4) charting and dictation; (5) contact with other providers; (6) patient-specific contact with housestaff; and (7) scheduling activities. MDs also estimated amount of time related to visit, but performed before or after the day of the visit, for 5 categories of EM services: (1) review of records; (2) talking to patient and family; (3) charting and dictation; (4) contact with other providers; and (5) scheduling and obtaining test results. Finally, MDs estimated total work for each visit and the proportion of total work performed on the day of the visit. Survey produced data on 19,143 visits.

2.	Develop new CPT codes for EM services	2.	a.	Convene 46-member consensus panel consisting of 29 MDs, 5 reps from CPT Editorial Panel, 8 reps from Medicare carriers and private insurers, 2 consumer reps, 1 nurse, and 1 physician assistant, to examine the following issues: (1) relationship between different measures of time and total work; (2) differences in MD practice styles and use of nonphysician providers; (3) specific EM services provided during visits of a particular duration; (4) impact of specific variables on the relationship between work and time; and (5) common encounter times for different classes of visits. Information was collected from panelists via telephone and mail surveys, as well as face-to-face meetings. Panel used data from Phase I, PPRC Survey of Visits and Consultations (see c, below), and AMA Ad Hoc Committee on Visits and Levels of Service. Panel recommended that EM codes recognize 3 classes of visit (new patient/initial care; established patient/subsequent care; and consultation) and 5 levels of service within each class based on content and "typical encounter time."
			b.	Convene additional panels to revise specialized EM services, such as psychiatric, ophthalmologic, ER, and preventive.
			c.	Refine recommendations of consensus panel into a visit coding system, consisting of 12 classes of visits and 5 levels of services within each class.
3.	Validate estimates of pre- and post-service work for surgical global services	3.	a.	Separate surgical global service into: (1) pre-op visits, (2) operative (including scrub work); and (3) post-op visits.
			b.	Use intra-service work and scrub work from Hsiao as measure of work for the operative component.
			c.	Obtain estimates of pre- and post-operative visits from specialty societies. Societies used either a consensus process or committee process. Methods to be documented in reports submitted to PPRC.
			d.	Validate specialty society data using claims data from carriers that do not include visits in the global fee, data from HMOs and multispecialty groups, and physician survey data. Also, convene a panel of MDs "not directly affected by payment reform" to assess the face validity of estimates for services where the objective data for comparison are inadequate.

		e.	Conduct survey of 56 carriers to obtain components of the surgical global package for 4 commonly performed procedures: (1) total hip replacement; (2) TURP; (3) CABG; and (4) permanent pacemaker insertion.
4.	Multi-step review and refinement of Hsiao's relative work values	4. a.	Refine RWVs for surveyed non-EM services using specialty societies. Consult with Harvard, AMA, and HCFA before assigning new RWVs. For disputed RWVs where data are inadequate, refer to a multi-specialty advisory panel.
		b.	Refine cross-specialty links. Use specialty societies to review all linkages. Consult with Harvard, AMA, and HCFA before suggesting alternative links.
		c.	Final refinement of common scale of work, using refined RWVs from specialty-specific panels. Assign RWVs for non-EM services performed by multiple specialties based on volume-weighted average. Apply a global adjustment to maintain budget neutrality within each category of service for each specialty.
		d.	Refine RWVs for imaging services.

	CPT Editorial Panel		Research Methods
1.	Develop new CPT codes for EM services	1.	Develop 6 categories of EM codes: (1) office/outpatient visit for new patients; (2) office/outpatient visits for established patients; (3) initial inpatient hospital care; (4) subsequent inpatient hospital care; (5) initial consultations; and (6) follow-up consultations. Each category is also divided into 3-5 levels of care, including measure of time.

	Specialty Societies		Research Methods
1.	Resurvey Phase I RWVs	1.	Society of Thoracic Surgeons developed its own RWVs under contract with Abt Associates. This study addressed following limitations of Phase I study: (1) biased estimates of peri-service work due to use of regression models rather than direct estimates; (2) failure to assess MD "fitness to rate" surveyed services; (3) lack of subspecialty representation; (4) inappropriate choice of standard service; (5) absence of important services from survey; (6) use of imprecise vignettes; and (7) biased estimates of total work due to charge-based extrapolation.

BIBLIOGRAPHY

American Medical Association, *Common Procedural Terminology, 4th Edition (CPT-4),* American Medical Association, Chicago, 1991.

Bankart, C. P., and S. W. Powers, "Individual Decisions and Group Consensus," *Journal of Social Psychology*, Vol. 126, 1986, pp. 369–374.

Becker, E. R., D. Dunn, and W. C. Hsiao, "Relative Cost Differences Among Physicians' Specialty Practices," *JAMA*, Vol. 260, No. 16, October 1988, pp. 2397–2402.

Berry, S., and D. Kanouse, "Physician Response to a Mailed Survey: An Experiment in Timing of Payment," *Public Opinion Quarterly*, Vol. 51, 1987, pp. 102–114.

Bock, R. D., and L. V. Jones, *The Measurement and Prediction of Judgment and Choice*, Holden-Day, San Francisco, 1968.

Braun, P., W. C. Hsiao, E. Becker, and M. DeNicola, "Evaluation and Management Services in the Resource-Based Relative Value Scale," *JAMA*, Vol. 260, No. 16, October 1988a, pp. 2409–2417.

Braun, P., D. B. Yntema, D. Dunn, M. DeNicola, T. Ketcham, D. Verrilli, and W. C. Hsiao, "Cross-Specialty Linkage of Resource-Based Relative Value Scales: Linking Specialties by Services and Procedures of Equal Work," *JAMA*, Vol. 260, No. 16, October 1988b, pp. 2390–2396.

Burton, G. E., "The 'Clustering Effect': An Idea-Generation Phenomenon During Nominal Grouping," *Small Group Behavior*, Vol. 18, 1987, pp. 224–238.

Dalkey, N. C., B. Brown, and S. W. Cochran, *The Delphi Method, III: Use of Self-Ratings To Improve Group Estimates*, RAND, RM-6115-PR, November 1969.

Davis, J. H., T. Kameda, C. Parks, M. Stasson, and S. Zimmerman, "Some Social Mechanics of Groups: The Distribution of Opinion, Polling Sequence, and Implications for Consensus," *Journal of Personality and Social Psychology*, Vol. 57, 1989, pp. 1000–1012.

Davis, J. H., R. S. Tindale, D. H. Nagao, V. B. Hinsz, and B. Robertson, "Order Effects in Multiple Decisions by Groups:

A Demonstration with Mock Juries and Trial Procedures," *Journal of Personality and Social Psychology*, Vol. 47, 1984, pp. 1003–1012.

Dempster, A. P., N. Laird, and D. B. Rubin, "Maximum Likelihood from Incomplete Data Using the EM Algorithm (with Discussion)," *Journal of the Royal Statistical Society, Series B*, Vol. 39, 1977, pp. 1–38.

Dukerich, J. M., M. L. Nichols, D. R. Elm, and D. A. Vollrath, "Moral Reasoning in Groups: Leaders Make a Difference," *Human Relations*, Vol. 43, 1990, pp. 473–493.

Dunn, D., W. C. Hsiao, T. R. Ketcham, and P. Braun, "A Method for Estimating the Preservice and Postservice Work of Physicians' Services," *JAMA*, Vol. 260, No. 16, October 1988, pp. 2371–2378.

Gilbert, W. J., *Modern Algebra with Applications,* Wiley, New York, 1976.

Glisson, C., "The Group Versus the Individual as the Unit of Analysis in Small Group Research," in S. D. Rose and R. A. Feldman (eds.), *Research in Social Group Work*, Haworth Press, New York, 1987, pp. 13–21.

Health Care Financing Administration, "Medicare Program; Fee Schedule for Physicians' Services; Proposed Rule," *Federal Register*, Vol. 56, No. 108, 5 June 1991a, pp. 25792–25978.

Health Care Financing Administration, "Medicare Program; Fee Schedule for Physicians' Services," *Federal Register*, Vol. 56, No 227, 25 November 1991b, pp. 59501–59819.

Hinsz, V. B., D. A. Vollrath, D. H. Nagao, and J. H. Davis, "Comparing the Structure of Individual and Small Group Perceptions," *International Journal of Small Group Research*, Vol. 4, 1988, pp. 159–168.

Hsiao, W. C., P. Braun, E. R. Becker, D. L. Dunn, N. L. Kelly, and D. B. Yntema, *A National Study of Resource-Based Relative Value Scales for Physician Services: Phase II. Final Report*, Harvard School of Public Health, Cambridge, Massachusetts, 1990.

Hsiao, W. C., P. Braun, D. Dunn, and E. R. Becker, "Resource-Based Relative Values: An Overview," *JAMA*, Vol. 260, No. 16, October 1988a, pp. 2347–2353.

Hsiao, W. C., P. Braun, N. L. Kelly, and E. R. Becker, "Results, Potential Effects, and Implementation Issues of the Resource-

Based Relative Value Scale," *JAMA*, Vol. 260, No. 16, October 1988b, pp. 2429–2438.

Hsiao, W. C., N. P. Couch, N. Causino, E. R. Becker, T. R. Ketcham, and D. K. Verrilli, "Resource-Based Relative Values for Invasive Procedures Performed by Eight Surgical Specialties," *JAMA*, Vol. 260, No. 16, October 1988c, pp. 2418–2424.

Hsiao, W. C., D. B. Yntema, P. Braun, D. Dunn, and C. Spencer, "Measurement and Analysis of Intraservice Work," *JAMA*, Vol. 260, No. 16, October 1988d, pp. 2361–2370.

Irwin, D., P. Lutz, S. F. Moore, and L. S. Shaffer, "Group Versus Individual Decision Making in the Commons Problem," *Journal of Social Psychology*, Vol. 129, 1988, pp. 551–553.

Janis, I. L., *Victims of Groupthink*, Houghton Mifflin, Boston, Massachusetts, 1972.

Kahan, J. P., *Indirect and Direct Ratio Scaling Methods for Diverse Psychological Constructs*, Ph.D. Dissertation, the University of North Carolina, Chapel Hill, North Carolina, 1968.

Kahneman, D., P. Slovic, and A. Tversky, *Judgment Under Uncertainty: Heuristics and Biases*, Cambridge University Press, Cambridge, Massachusetts, 1982.

Kelly, N. L., W. C. Hsiao, P. Braun, A. Sobol, and M. DeNicola, "Extrapolation of Measures of Work for Surveyed Services to Other Services," *JAMA*, Vol. 260, No. 16, October 1988, pp. 2379–2384.

Kerr, N. L., and J. Y. Huang, "Jury Verdicts: How Much Difference Does One Juror Make?" *Personality and Social Psychology Bulletin*, Vol. 12, 1986, pp. 325–343.

Lasker, R., M. S. Marquis, and M. Morrow, *Survey of Visits and Consultations,* Physician Payment Review Commission, Report No. 91-1, Washington, D.C., 1991.

Laughlin, P. R., "Social Combination Processes of Cooperative Problem-Solving Groups on Verbal Intellective Tasks," in M. Fishbein (ed.), *Progress in Social Psychology, Vol. 1*, Lawrence Erlbaum Associates, Hillsdale, New Jersey, 1980.

Laughlin, P. R., and G. C. Futoran, "Collective Induction: Social Combination and Sequential Transition," *Journal of Personality and Social Psychology*, Vol. 48, 1985, pp. 608–613.

Laughlin, P. R., and R. P. McGlynn, "Collective Induction: Mutual Group and Individual Influence by Exchange of Hypotheses and

Evidence," *Journal of Experimental Social Psychology*, Vol. 22, 1986, pp. 567–589.

Lee, P. R., and P. B. Ginsburg, "Physician Payment Reform: An Idea Whose Time Has Come," *JAMA*, Vol. 260, No. 16, October 1988, pp. 2441–2443.

Little, R. J. A., and D. B. Rubin, *Statistical Analysis with Missing Data*, Wiley, New York, 1987.

Maloney, J. V., "A Critical Analysis of the Resource-Based Relative Value Scale," *JAMA*, Vol. 266, No. 24, December 1991, pp. 3453–3458.

McGrath, J. E., *Groups: Interaction and Performance*, Prentice-Hall, Englewood Cliffs, New Jersey, 1984.

Meyers, R. A., "Persuasive Arguments Theory: A Test of Assumptions," *Human Communication Research*, Vol. 15, 1989a, pp. 357–381.

Meyers, R. A., "Testing Persuasive Argument Theory's Predictor Model: Alternative Interactional Accounts of Group Argument and Influence," *Communication Monographs*, Vol. 56, 1989b, pp. 112–132.

Michaelson, L. K., W. E. Watson, and R. H. Black, "A Realistic Test of Individual Versus Group Consensus," *Journal of Applied Psychology*, Vol. 74, 1989, pp. 834–839.

Mosteller, F., and J. W. Tukey, *Data Analysis and Regression*, Addison-Wesley, Reading, Massachusetts, 1977.

Nichols, M. L., and V. E. Day, "A Comparison of Moral Reasoning of Groups and Individuals on the 'Defining Issues Test,'" *Academy of Management Journal*, Vol. 25, 1982, pp. 201–208.

Noether, M., D. Wierz, M. Hecker, and H. Goldberg, *Development of a Resource-Based Relative Value Scale for Cardiothoracic and Vascular Surgery*, Abt Associates, Cambridge, Massachusetts, 1990.

Ono, K., and J. H. Davis, "Individual Judgment and Group Interaction: A Variable Perspective Approach," *Organizational Behavior and Human Decision Processes*, Vol. 41, 1988, pp. 211–232.

Pasnak, R., "Evaluation of a National Study of Resource-Based Relative Value Scales for Physician Services," paper prepared for the American College of Surgeons, n.d.

Physician Payment Review Commission, *Annual Report to Congress, 1989*, Washington, D.C., 1989.

Physician Payment Review Commission, *Annual Report to Congress, 1991*, Washington, D.C., 1991.

Prentice-Dunn, S., and R. W. Rogers, "Deindividuation and the Self-Regulation of Behavior," in P. B. Paulus (ed.), *Psychology of Group Influence, 2nd Edition*, Lawrence Erlbaum Associates, Hillsdale, New Jersey, 1989.

Roper, W. L., "The Resource-Based Relative Value Scale: A Methodological and Policy Evaluation," *JAMA*, Vol. 260, No. 16, October 1988, pp. 2444–2446.

Stasser, G., N. L. Kerr, and J. H. Davis, "Influence Processes and Consensus Models in Decision-Making Groups," in P. B. Paulus (ed.), *Psychology of Group Influence, 2nd Edition*, Lawrence Erlbaum Associates, Hillsdale, New Jersey, 1989.

Stasser, G., and W. Titus, "Effects of Information Load and Percentage of Shared Information on the Dissemination of Unshared Information During Group Discussion," *Journal of Personality and Social Psychology*, Vol. 53, 1987, pp. 81–93.

Stasson, M. F., and J. H. Davis, "The Relative Effects of the Number of Arguments, Number of Argument Sources and Number of Opinion Positions in Group-Mediated Opinion Change," *British Journal of Social Psychology*, Vol. 28, 1989, pp. 251–262.

Stasson, M. F., K. Ono, S. K. Zimmerman, and J. H. Davis, "Group Consensus Processes on Cognitive Bias Tasks: A Social Decision Scheme Approach," *Japanese Psychological Research*, Vol. 30, 1988, pp. 68–77.

Stephenson, G. M., N. K. Clark, and G. S. Wade, "Meetings Make Evidence? An Experimental Study of Collaborative and Individual Recall of a Simulated Police Interrogation," *Journal of Personality and Social Psychology*, Vol. 50, 1986, pp. 1113–1122.

Stevens, S. S., "On the Psychophysical Law," *Psychological Review*, Vol. 64, 1957, pp. 153–181.

Stevens, S. S., "A Metric for the Social Consensus," *Science*, Vol. 151, 1966, pp. 530–541.

Stevens, S. S., and E. H. Galanter, "Ratio Scales and Category Scales for a Dozen Continua," *Journal of Experimental Psychology*, Vol. 54, 1957, pp. 377–411.

Tindale, R. S., "Group vs Individual Information Processing: The Effects of Outcome Feedback on Decision Making," *Organizational Behavior and Human Decision Processes*, Vol. 44, 1989, pp. 454–473.

Tindale, R. S., J. H. Davis, D. A. Vollrath, D. H. Nagao, and V. B. Hinsz, "Asymmetrical Social Influence in Freely Interacting Groups: A Test of Three Models," *Journal of Personality and Social Psychology*, Vol. 58, 1990, pp. 438–449.

Todd, J. S., "At Last, A Rational Way to Pay for Physicians' Services?" *JAMA*, Vol. 260, No. 16, October 1988, pp. 2439–2341.

Turner, J. C., M. S. Wetherell, and M. A. Hogg, "Referent Informational Influence and Group Polarization," *British Journal of Social Psychology*, Vol. 28, 1989, pp. 135–147.

Vollrath, D. A., B. H. Sheppard, V. B. Hinsz, and J. H. Davis, "Memory Performance by Decision-Making Groups and Individuals," *Organizational Behavior and Human Decision Processes*, Vol. 43, 1989, pp. 289–300.